Application of biclustering algorithm to extract rules from labeled data

Zhang Yanjie and Sun Hongbo
School of Computer and Control Engineering, Yantai University, Yantai, China

Abstract
Purpose – For many pattern recognition problems, the relation between the sample vectors and the class labels are known during the data acquisition procedure. However, how to find the useful rules or knowledge hidden in the data is very important and challengeable. Rule extraction methods are very useful in mining the important and heuristic knowledge hidden in the original high-dimensional data. It can help us to construct predictive models with few attributes of the data so as to provide valuable model interpretability and less training times.

Design/methodology/approach – In this paper, a novel rule extraction method with the application of biclustering algorithm is proposed.

Findings – To choose the most significant biclusters from the huge number of detected biclusters, a specially modified information entropy calculation method is also provided. It will be shown that all of the important knowledge is in practice hidden in these biclusters.

Originality/value – The novelty of the new method lies in the detected biclusters can be conveniently translated into if-then rules. It provides an intuitively explainable and comprehensive approach to extract rules from high-dimensional data while keeping high classification accuracy.

Keywords Biclustering algorithm, Crowdsourced big data and analytics, Rule extraction

Paper type Research paper

1. Introduction
Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product. These products are often proteins, but in non-protein coding genes such as ribosomal RNA, transfer RNA or small nuclear RNA genes, the product is a functional RNA. Gene expression data is a kind of data matrix used to represent the expression level of different genes under specific conditions simultaneously. Each element is a real number which is often the logarithm of the relative abundance of the mRNA of the gene (Madeira and Oliveira, 2004). Usually, in the data matrix, genes are arranged in the row direction, while the column direction represents different time or different environmental conditions.
As an array contains tens of thousands of probes, a microarray experiment can accomplish many genetic tests in parallel. The acquired gene expression data are a typical kind of high-dimensional data. The huge number of features or attributes adds great difficulties to the prediction and interpretability capabilities of all of the models applied to analyze it because of redundant features and noises. How to find the useful rules or knowledge hidden in the data is very important and challenging. Feature selection is usually a necessary step to facilitate further processing which is especially true for high dimensional data (Gorzalczy and Rudziński, 2017). In machine learning and statistics, attribute selection is the process of selecting a subset of the fewest number of informative attributes for classification, rule extraction and the other applications (Shrivastava and Barua, 2015).

Rule-based expert systems are often applied to classification problems in various application fields, like fault detection, biology and medicine (Dahal et al., 2015; Shrivastava and Barua, 2015). In Roubos et al.’s (2003) study, the authors show a compact, accurate and interpretable fuzzy rule-based classifiers obtained from labeled observation data. To implement it, an iterative approach for developing fuzzy classifiers was proposed. The initial model was derived from the data and subsequently, feature selection and rule-base simplification were applied to reduce the model, while a genetic algorithm was used for parameter optimization. Moreover, the researchers proposed different optimization-based methods such as ant colony optimization and particle swarm optimization to extract rules (Chen et al., 2015; Indira and Kanmani, 2015).

Support vector machines (SVMs) are learning systems based on the statistical learning theory and exhibit good generalization ability on different kinds of real data sets (Han et al., 2015). Companying with the study on SVMs, it has gradually turned into a leading machine learning technique and has been applied in a wide range of areas such as bioinformatics, pattern recognition, text classification and so on (Shi et al., 2015). Researchers interested in this topic can easily access to a lot of free software or toolbox. However, the results given by SVMs are usually difficult to explain. In safety-critical or medical applications, an explanation capability is an absolute requirement. A rule extraction method based on SVMs was proposed in Núñez et al.’s (2002) study. The authors introduced a SVM plus prototypes procedure for rule extraction. This method allows giving explanation ability to SVMs. Once determined the decision function by means of a SVM, a clustering algorithm was used to determine prototype vectors for each class. These points were combined with the support vectors using geometric methods to define ellipsoids in the input space with minimum overlapping between classes, which were later transferred to if-then rules.

One important analysis task of microarray data concerns the simultaneous identification of groups of genes that show similar expression patterns across specific groups of experimental conditions (Wang et al., 2014; Maulik et al., 2015). Most of the time, it is not the sample vectors as integrity shows the strong coherence with each other, but the elements at some specific positions among different sample vectors show the local similarity (Valarmathi et al., 2015). Besides classical clustering methods such as hierarchical clustering, in recent years, biclustering has become a popular approach to analyze biological data sets and a wide variety of algorithms, and analysis methods have been published (Czibula et al., 2015; Shinde and Kulkarni, 2016; Indira and Kanmani, 2015).

Such applications can be addressed by a biclustering process whose aim is to discover biclusters (Cheng and Church, 2000). The so called bicluster is a subset of genes and conditions of the original expression matrix where the selected genes present a coherent
behavior under all the experimental conditions contained in the bicluster. In other words, the data in the same bicluster show a high degree of local similarity. The difference between a bicluster and a submatrix is that all the biclusters are definitely submatrices, but only those submatrices whose row or column vectors satisfying some kind of linear relations will be treated as biclusters. Biclustering algorithms are just a kind of data processing algorithms to find those submatrices lying in the original data matrix showing the local similarity. This technology has found numerous applications in research and applied areas like biology, drug discovery, toxicological study and diseases diagnosis (Alon et al., 1999; Alizadeh et al., 2000; Golub et al., 1999; Pomeroy et al., 2002).

However, the number of biclusters lying in the data, the size and the spatial positioning relations among these biclusters is completely unknown and strongly data dependent (Rabia et al., 2016). In Kaiser and Leisch’s (2008) study, the authors introduced the R package which contains a collection of biclustering algorithms, preprocessing methods for two way data and validation and visualization techniques for bicluster results. In Amela et al.’s (2006) study, the authors provided the Biclustering Analysis Toolbox, BicAT, as a software platform for clustering-based data analysis that integrates various biclustering and clustering techniques in terms of a common graphical user interface. Furthermore, BicAT provides different facilities for data preparation, inspection and post processing such as discretization, filtering of biclusters according to specific criteria or gene pair analysis for constructing gene interconnection graphs. The toolbox is described in the context of gene expression analysis but is also applicable to other types of data. The authors compared different biclustering techniques with each other with respect to the biological relevance of the clusters as well as with other characteristics such as robustness and sensitivity to noise (Shi et al., 2015; Maulik et al., 2015).

When the biclusters have been detected by applying the biclustering algorithm, the problem is how to translate the biclusters into the corresponding rules. In fact, it can be easily implemented combining with the data discretization schemes. As each bicluster is a submatrix, the line and column numbers that it covers are known. As each experiment condition can be treated as an attribute and all the column numbers of the bicluster can be used as a prerequisite for a rule, a bicluster detection procedure can also be thought as an attribute’s selection processing. This kind of rule extraction provides a comprehensive interpretable way compared with the other methods while keeping high classification accuracy.

2. The proposed method

For many pattern recognition problems, the relation between the samples and the classes are known during the acquisition procedure of the data. And this kind of data is called labeled data. Suppose a bicluster \( B \) is composed of the row numbers set \( \{i_1, i_2, \ldots, i_m\} \) and column numbers set \( \{j_1, j_2, \ldots, j_n\} \) of a labeled data matrix \( D \), then the function \( \Phi(B) = \{i_1, i_2, \ldots, i_m\} \) is defined to determine the set of those row numbers that the elements of \( B \) lies in \( D \) and \( |\Phi(B)| = m \), so is the definition of the function \( \Psi(B) = \{j_1, j_2, \ldots, j_n\} \). The position of a bicluster \( B \) lying in the original data matrix \( D \) can be determined by \( \Phi(B) \) together with \( \Psi(B) \).

Usually, the data in the same class show some kind of behavior similarity is called a rule. A rule is applicable only for the sample vectors in the same class. The similarity of sample vectors spanning over the class boundary cannot be thought of the knowledge to distinguish data among different classes. That means biclustering processing results with labeled data are meaningful depending on the class labels. Directly biclustering with \( D \) without
considering the labels of the sample vectors does not help to find the biclusters which will be translated into the rules eventually.

2.1 Data discretization

The flowchart of the new rule extraction method is illustrated in Figure 1, which is mainly composed of four sequential processing procedures: data discretization, biclustering processing, bicluster significance evaluation and rule translation based on the discretization schemes. Data discretization is a technique to partition continuous attributes into a finite set of adjacent intervals to generate attributes with a small number of distinct values (Kurgan and Cios, 2004). Discretization algorithms have played an important role in data mining and knowledge discovery (Tsai et al., 2008). They not only produce a concise summarization of continuous attributes to help the experts understand the data more easily but also make learning more accurate and faster (Oliveira, 1999).

Assuming that a dataset consists of \( M \) examples and \( S \) target classes, a discretization algorithm would discretize the continuous attribute \( a \) in this dataset into \( n \) discrete intervals \( \{[d_0, d_1], [d_1, d_2], \ldots, [d_{n-1}, d_n]\} \), where \( d_0 \) is the minimal value and \( d_n \) is the maximal value of attribute \( a \). Such a discrete result is called a discretization scheme on attribute \( a \). This discretization scheme should keep the high interdependency between the discrete attribute and the target class to carefully avoid changing the distribution of the original data.

As having been introduced before, each column of \( D \) can be considered as an attribute or feature no matter what real physical meaning it has. If a sample vector \( V \in D \) has value \( V(a) \) with respect to the attribute \( a \), then the discretized value \( V_D(a) \) of \( V \) on \( a \) is determined by the discretization scheme. For an example, if it is known that \( V(a) \in (d_i, d_{i+1}) \), \( i = 0, 1, \ldots, n - 1 \), then after the discretization processing, the value of \( V_D(a) \) will be \( i + 1 \).

The wine data contains the chemical analysis of 178 wines produced in the same region in Italy but derived from three different cultivars. The problem is to distinguish the three different types based on 13 continuous attributes derived from chemical analysis: alcohol, malic acid, ash, alcalinity of ash, magnesium, total phenols, flavanoids, nonflavanoids phenols, proanthocyaninsm color intensity, hue, OD280/OD315 of diluted wines and proline (Roubos et al., 2003). In Figure 2, the original wine data with all of its 13 attributes are shown. The wine data are also a kind of high-dimensional data even though their dimensionality is less than the real gene expression data. It has been widely applied in the research studies on machine learning such as attribute selection, pattern recognition and rule extraction. The discretized wine data with all of their 13 attributes are shown in Figure 3. The data discretization schemes are listed in Table I where the method proposed in Tsai et al.’s (2008) study is applied.
Figure 2.
The original wine data with all of the 13 attributes.
Figure 3. The discretized wine data with all of the 13 attributes.
2.2 Criterion for rule extraction

The symbol $B_j|\omega$, $j = 1, 2, \ldots, m$ is used to represent all of the biclusters detected within the sample vectors belonging to the same class $\omega$. Then $|\Phi(B_j|\omega)|$ and $|\Psi(B_j|\omega)|$ are two very important factors which can be used to evaluate the significance of the bicluster $B_j|\omega$. As we expect the number of rules should be as small as possible, the requirement has the meaning in twofold. First, the smallest number of rules means the nature of the data has been well grasped by the rules. Second, the rules should have restriction on all of the vectors within the same class $\omega$. In other words, the rules should be adapted to all of the vectors in the same class. The above analysis tells us the first rule selection criterion which is:

$$|\Phi(B_j|\omega)| + |\Phi(B_{j+1}|\omega)| + \cdots + |\Phi(B_m|\omega)| \geq |\Phi(\omega)|$$

In equation (1), the subscript variable $m$ means the number of biclusters within the same class $\omega$ has been used to extract the corresponding rules. Undoubtedly, the value of $m$ should be as small as possible.

Usually, there are a huge number of biclusters which can be detected. The significance of each detected bicluster must be evaluated by calculating its information entropy. Based on it, the significance of all these biclusters can be sorted in a decreased manner. Applying the first rule selection criteria, the minimum number $m$ can be determined. It must be pointed out that since every bicluster is corresponded to a rule, then minimum number of rules to well express the knowledge hidden in the class $\omega$ is also $m$. Here, we simultaneously draw an important conclusion which is the least number of rules providing 100 per cent of recognition accuracy is data-dependent. We also give the way to determine the exact value of it. However, when the number of rules is fixed, the new algorithm provides the most distinct and convenient way to find the rules while giving the assurance of the maximum accuracy.

Though the introduced above rule extraction method from bicluster is done within each of the class respectively. Whether the combination of rules from different classes can be applied to the whole data should be done with careful consideration. If each column of the data $D$ is treated as an attribute, then it usually has different values. The information entropy corresponding to each attribute within the same class $\omega$ can be calculated as

**Table I. Discretization schemes of all the attributes of the wine data**

<table>
<thead>
<tr>
<th>No.</th>
<th>Attribute name</th>
<th>$d_0$</th>
<th>$d_1$</th>
<th>$d_2$</th>
<th>$d_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alcohol</td>
<td>11.030</td>
<td>12.780</td>
<td>14.830</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Malic acid</td>
<td>0.740</td>
<td>1.475</td>
<td>2.235</td>
<td>5.800</td>
</tr>
<tr>
<td>3</td>
<td>Ash</td>
<td>1.360</td>
<td>2.030</td>
<td>3.230</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Alcalinity of ash</td>
<td>10.600</td>
<td>17.900</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Magnesium</td>
<td>70</td>
<td>88.500</td>
<td>162</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Total phenols</td>
<td>0.980</td>
<td>1.840</td>
<td>2.335</td>
<td>3.880</td>
</tr>
<tr>
<td>7</td>
<td>Flavanoids</td>
<td>0.340</td>
<td>1.400</td>
<td>2.310</td>
<td>5.080</td>
</tr>
<tr>
<td>8</td>
<td>Nonflavanoid phenols</td>
<td>0.130</td>
<td>0.395</td>
<td>0.660</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Proanthocyanins</td>
<td>0.410</td>
<td>1.205</td>
<td>1.655</td>
<td>3.580</td>
</tr>
<tr>
<td>10</td>
<td>Color intensity</td>
<td>1.280</td>
<td>3.820</td>
<td>7.550</td>
<td>13</td>
</tr>
<tr>
<td>11</td>
<td>Hue</td>
<td>0.480</td>
<td>0.785</td>
<td>1.710</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>OD280/OD315</td>
<td>1.270</td>
<td>2.115</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Proline</td>
<td>278</td>
<td>735</td>
<td>1680</td>
<td></td>
</tr>
</tbody>
</table>
\[ H(a|\omega) = \sum_{i=1}^{N(a|\omega)} -p_i \ln p_i, \]  
the variable \( N(a|\omega) \) means the number of different values corresponding to the attribute \( a \) in the class \( \omega \). Suppose that all of the sample vectors in \( D \) can be classified into \( n \) classes, if \( H(a|\omega) = \frac{1}{n} H(a|\omega_i) \) then the attribute \( a \) is of the least importance, as all of the sample vectors have the same value on it.

2.3 Bicluster significance evaluation

How to select the significant bicluster among the huge number of detected biclusters is very important. Given a data matrix \( D \), each row vector of \( D \) can be considered as a sample vector, assume the number of sample vectors lying in \( D \) is \( N \), all these samples belong to different classes named \( \omega_i \), \( i = 1, 2, \ldots, M \). Define \( N_i \) is the number of samples in class \( \omega_i \), then \( p_i \) which means the probability of one sample belonging to the class \( \omega_i \) can be estimated by \( \frac{N_i}{N} \). The expected information entropy provided by \( D \) is:

\[ I(N_1, N_2, \ldots, N_M) = -\sum_{i=1}^{M} p_i \log_2 p_i \]  
(2)

If an attribute \( A \) has a number of \( k \) values which are \{a_1, a_2, \ldots, a_k\}, then the whole samples set can be classified into \( k \) different subset \( S_1, S_2, \ldots, S_k \) by only using attribute \( A \). Assume \( N_{ij} \) is the number of samples in the subset \( S_j \) which belongs to the class \( \omega_i \), then the information entropy of classified result by attribute \( A \) is defined as:

\[ I(A) = \sum_{j=1}^{k} \left[ \frac{N_0 + N_1 + \cdots + N_k}{N} I(N_{1j}, N_{2j}, \ldots, N_{Mj}) \right] \]  
(3)

where \( I(N_{1j}, N_{2j}, \ldots, N_{Mj}) = -\sum_{i=1}^{M} p_{ij} \log_2 p_{ij} \) and \( p_{ij} = \frac{N_{ij}}{N_j} \) which means the probability of samples lying in subset \( S_j \) belonging to the class \( \omega_i \). The whole information gain acquired by attribute \( A \) is:

\[ Gain(A) = I(N_1, N_2, \ldots, N_M) - I(A) \]  
(4)

Assume that there is a bicluster \( B \) when doing biclustering with \( E(D) \), it is a fact that the more row numbers \( B \) covers and the less attributes \( B \) has, the more meritorious \( B \) is. Based on it, we define \( \Delta(B) = \frac{|\Phi(\omega)|}{|\Psi(B)|} \) as a weight to indicate the importance of information provided by \( B \). \( |\Phi(\omega)| \) is the number of samples belonging to the class \( \omega \) where the bicluster \( B \) is founded. Equation (3) only instructs how to calculate the information entropy with one attribute. As each bicluster \( B \) satisfies \( |\Psi(B)| \geq 2 \), which means we have to take a number of \( |\Psi(B)| \) attributes’ information entropy into account. For any two different attributes, \( A_1 \) and \( A_2 \) if \( I(A_1) < I(A_2) \), then the values taken by \( A_1 \) are more regular than the values taken by \( A_2 \). When applying these two attributes to classify an unknown input sample, the classified result based on \( A_1 \) will be more accurate than that of \( A_2 \). Considering all of
Among all these biclusters, 217 biclusters are within the aforementioned analysis, we define the following formula as an index to evaluate the significance of the bicluster $B$:

$$\text{IE}(B) = \Delta(B) \max_{A \in |\Psi(B)|} \text{Gain}(A)$$  \hspace{1cm} (5)

### 2.4 Translation of biclusters to rules

Assume there is a bicluster $B|\bar{\omega}$ lying in the class $\bar{\omega}$ where the number of sample vectors in $\bar{\omega}$ is $\Phi(\bar{\omega}) = N \bar{\omega}$ satisfying $\Phi(B|\bar{\omega}) = \{i_1, i_2, \cdots, i_m\}$, $\Psi(B|\bar{\omega}) = \{j_1, j_2, \cdots, j_n\}$, $m \leq N \bar{\omega}$, $n \leq \Psi(D)$, the bicluster $B|\bar{\omega}$ can be conveniently translated into the corresponding rule accompanied by the data discretization schemes. As there are a number of $|\Psi(B|\bar{\omega})| = n$ attributes in $B|\bar{\omega}$, the translated rule has $n$ antecedents which are related with attributes $a_{j_1}, a_{j_2}, \cdots, a_{j_n}$, respectively. Here, the attribute $a_{j_i}$ is used for explanation. If the data discretization scheme on attribute $a_{j_i}$ is $\{[d_0, d_1], (d_1, d_2), \cdots, (d_{n-1}, d_n]\}_{a_{j_i}}$, as we have known how the data discretization works, the value $j_1$ on which the attribute $a_{j_i}$ is means the original attribute $a_{j_i}$’s value without doing data discretization belongs to the range $(d_{j_i-1}, d_{j_i}]$, by this way, the first antecedent of the rule is if $a_{j_i} \in (d_{j_i-1}, d_{j_i}]$. Keep on this kind of processing till to the attribute $a_{j_n}$, the full description of the rule is determined.

### 3. Computation example

The well-known wine data are applied as the experiment data to illustrate the feasibility and effectiveness of the proposed new method. There are 59 samples in class $\omega_1$, 71 samples in class $\omega_2$ and 48 samples in class $\omega_3$. The number of all of the samples is 178. Each sample vector has 13 numerical attributes whose values are different from each other observably. Compared with the real gene expression data, the wine data have smaller dimensionality, while they are all numerical data with high dimensionality (Wang et al., 2014). The application of wine data will help to save a lot of time to verify the feasibility of the proposed algorithm without destroying the nature of the research data object. And it also facilitates the comparison among the different research results.

Here, the data discretization method proposed in Tsai et al.’s (2008) study is applied and the discretization schemes are listed in Table I. The discretized wine data are illustrated in Figure 4. As there are three classes, the discretized data within each class is isolated as a single picture. According to the data discretization scheme, each of the 13 attributes is discretized into less than three intervals which means the discretized data are only composed of three different numbers 1, 2 and 3. Each number is represented by a colorful square for intuitive illustration. The whole processing is followed by the procedures shown in Figure 1.

Biclustering with the discretized wine data, a number of 5,716 biclusters are founded. Among all these biclusters, 217 biclusters are within $\omega_1$, 3,794 biclusters are within $\omega_2$ and 1,705 biclusters are within $\omega_3$. Using equation (5) to calculate every biclusters’ IE, biclusters $B_1$, $B_2$ and $B_3$ belonging to the class $\omega_1$, $\omega_2$ and $\omega_3$ are selected respectively. Accompanying with the data discretization schemes listed in Table 1, the selected biclusters can be translated into three rules as follows:

1. $\Psi(B_1) = \{1, 3, 5, 11, 12, \omega_1\}$, $|\Phi(B_1)| = 59$, $\frac{|\Phi|}{|\Psi|} = 100\%$, each row vector of $B_1$ is $\{2, 2, 2, 2, 2\}$, $B_1$ is translated into the rule as: for a sample vector $\bar{\alpha}$, if $a_1 \in (12.78, 14.83]$ and $a_3 \in (2.03, 3.23]$ and $a_5 \in (88.5, 162]$ and $a_{11} \in (0.785, 1.71]$ and $a_{12} \in (2.115, 4]$ then $\bar{\alpha} \in \omega_1$. 

2. 

3. 


(2) \( \Psi(B_2) = \{13, \omega_2\}, |\Phi(B_2)| = 67 \), \( \frac{|\Phi(B_2)|}{\Phi(\omega_2)} = 94.37\% \), each row vector of \( B_2 \) is (13, 2), \( B_2 \) is translated into the rule as: for a sample vector \( \alpha \), if \( a_{13} \in (278, 755) \) then \( \alpha \in \omega_2 \).

(3) \( \Psi(B_3) = \{3, 4, \omega_3\}, |\Phi(B_3)| = 47 \), \( \frac{|\Phi(B_3)|}{\Phi(\omega_3)} = 97.92\% \), each row vector of \( B_3 \) is (2, 2, 3), \( B_3 \) is translated into the rule as: for a sample vector \( \alpha \), if \( a_3 \in (2.03, 3.23) \) and \( a_4 \in (17.9, 30) \) then \( \alpha \in \omega_3 \).

As these three biclusters totally cover 173 sample vectors out of the whole 178 sample vectors, the three extracted knowledge offer a recognition accuracy of 97.19 per cent.

### 4. Conclusions

Rule extraction methods as an approach which tries to find the useful knowledge hidden in the high-dimensional data are very useful. The so-called rules are in practice, and some sample vectors in the data show coherent similarity with each other. Because the data in the same bicluster are closely related to each other, the transition from bicluster to rule has a natural consistency. As the elements of a bicluster are just lying in the original data and can be conveniently translated into a corresponding rule, the results of the new method have good explanation ability. The difference of the new method with the other methods lies in it applies the biclustering algorithm to discover the local similar biclusters existing among the original data matrix.

In the large amount of bicluster results detected, a specially modified information entropy calculation method is provided to evaluate the significance of all the detected biclusters. Then, all of the biclusters can be sorted in a decreased manner according to their information entropy values. By this way, those most significant biclusters within the samples belonging to the same class individually can be selected to extract the rules.
try to deal with more different types of data and compare the results with the results of existing literature. Processing with real gene expression data is ongoing and will be presented in the future work.

References


Further reading


About the authors

Zhang Yanjie was born in Wendeng, Shandong Province, China on November 16th, 1975. In 2004, he got his doctor’s degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, Beijing, China. In 2001, he got his master’s degree in information and control engineering, and in 1998, he got his bachelor’s degree in mathematics and systematic science from Shandong University, Jinan, China. He is an Associate Professor of Yantai University, Shandong, China now. His research interests include pattern recognition, bioinformatics and information security. Zhang Yanjie is the corresponding author and can be contacted at: yanjie.zhang@126.com
Sun Hongbo became a Member (M) of IEEE in 2010. Hongbo was born in Fuping, Shanxi Province, China, on 13th February, 1977. In 2011, he got his doctor's degree in control science from Tsinghua University, Beijing, China. In 2005, he got his master's degree in software engineering of Tsinghua University, and in 1998, he got his bachelor's degree in information science from Beijing Institute of Technology, China. He is a Lecturer of Yantai University, Shandong, China now. He has been a Postdoctoral Researcher in Department of Automation, Tsinghua University, Beijing, China, from July 2011 to June 2014. From November 2009 to November 2010, he has worked in National Research Council Canada as an International Visiting Worker. Between 2005 and 2006, he served as a Software Engineer of National CIMS ERC, Tsinghua University, Beijing, China. And during 1998 to 2002, he was an Assistant Professor of Shenyang Institute of Technology, Liaoning, China. His research interests include system integration, artificial intelligence, algorithm, large-scale simulation and e-commerce.

For instructions on how to order reprints of this article, please visit our website: www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: permissions@emeraldinsight.com
Quality-time-complexity universal intelligence measurement

Jing Liu, Zhiwen Pan, Jingce Xu, Bing Liang, Yiqiang Chen and Wen Ji

Beijing Key Laboratory of Mobile Computing and Pervasive Device,
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

Abstract

Purpose – With the development of machine learning techniques, the artificial intelligence systems such as crowd networks are becoming more autonomous and smart. Therefore, there is a growing demand for developing a universal intelligence measurement so that the intelligence of artificial intelligence systems can be evaluated. This paper aims to propose a more formalized and accurate machine intelligence measurement method.

Design/methodology/approach – This paper proposes a quality–time–complexity universal intelligence measurement method to measure the intelligence of agents.

Findings – By observing the interaction process between the agent and the environment, we abstract three major factors for intelligence measure as quality, time and complexity of environment.

Originality/value – This paper proposes a calculable universal intelligent measure method through considering more than two factors and the correlations between factors which are involved in an intelligent measurement.

Keywords Turing test, Agent–environment framework, Algorithmic information theory, Kolmogorov complexity, Universal intelligence

Paper type Research paper

1. Introduction

1.1 Background and related work
With the development of machine learning techniques, the artificial intelligence systems such as crowd networks are becoming more autonomous and smart. Therefore, to evaluate the intelligence of artificial intelligence systems, the universal intelligence measurement is needed. The current intelligence measurement methods can be classified as the human IQ test and the measurement of machine intelligence. IQ test is mainly through people’s perception of knowledge, text and graphics and understanding to test the intelligence of individuals. The machine intelligence can be measured based on human discrimination, problem benchmark, task response theory estimation and algorithmic information theory (Hernández-Orallo, 2014; Solomonoff, 2009).

© Jing Liu, Zhiwen Pan, Jingce Xu, Bing Liang, Yiqiang Chen and Wen Ji. Published in International Journal of Crowd Science. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at http://creativecommons.org/licenses/by/4.0/legalcode

This work is supported by the National Key Research & Development Plan of China (2017YFB1400100), the National Natural Science Foundation of China (61572466), and the Beijing Natural Science Foundation (4162059).
In a crowd network, a number of intelligent agents are able to collaborate with each other to finish a certain kind of sophisticated task (Prpic and Prashant, 2016). How to allocate the tasks to the agents in an optimized manner is the primary concern of a crowd network. As agents obtain different abilities (e.g. profession and reliability), the optimized task allocation should be performed on the basis of the evaluation of agents abilities. However, the agents are inherently heterogeneous for they operate within a hybrid-space including information space, physical space and awareness space and such hybrid-space varies with profession and tasks of corresponding agents. Therefore, it is not feasible to evaluate the ability of agent in a comprehensive manner.

Performing intelligence measurement on agents is one of the feasible ways to evaluate the ability of agents partially. The research of intelligent measurement can be dated back to 1950 when Turing test was proposed by Alan M. Turing (Turing, 1950). In recent years, a number of intelligent measurement methods have been proposed by Oppy and David (2003); Longo (2009); Mahoney (1999); Gibson (1998); Masum et al (2002); Alvarado et al (2001) and Smith (2006). However, according to the results of these papers, all the proposed methods have the following drawbacks:

- None of these methods are comprehensive enough to make the measurements by considering more than two factors including reward quality, timeliness, complexity of the environment, etc.
- Most of these methods (Oppy and David, 2003; Mahoney, 1999; Masum et al., 2002; Alvarado et al., 2001; Smith, 2006) do not evaluate the correlations between factors that are involved in an intelligent measurement. Hence, the effectiveness of the selected factors cannot be proved.

1.2 Summary of content and contributions
In this paper, we propose an intelligent measure approach for intelligent machines such as the agents in crowd network. We name the approach as QTC (quality–time–complexity) intelligent measure approach, as it can perform intelligent measurement by considering three factors: test complexity, rewards quality and timeliness. We proved that there are correlations between the reward quality and the two other factors. The intelligence of an intelligent agent is quantified through calculating the expected accumulative reward quality of the agent.

The rest of the paper is organized as follows. In Section 2, we briefly introduce the agent–environment framework for conducting the intelligent test and then introduce the three factors for intelligence measure. In Section 3, the correlations between the reward quality and the two other factors are evaluated. Then the QTC intelligent measure approach is introduced in detail. In Section 4, we prove the effectiveness of our approach by implementing a famous intelligent measure test. We conclude the paper in Section 5.

2. Agent–environment framework for conducting intelligent test
There are two steps for measuring the intelligence of an agent. The first step is conducting intelligent test on the agent so that the outcome of the test can be collected for further analysis. The second step is using intelligent measure approach to analyze the information collected from the intelligent test. In this paper, we conduct intelligent tests based on a widely accepted agent–environment framework. The detail of agent–environment framework and its implementation is introduced in this section. Agent–environment framework is a widely accepted framework that provides a guideline to conduct the intelligent test. As shown in Figure 1, there are three components in this framework: an
agent, an environment and a goal (Legg and Hutter, 2006). Agent is an intelligent entity who is taking the test. The goal is the task assigned to the agent during the test. The goal of a test is predefined by the test designer and should be informed to the agent before the test. The environment is a space to control the agent; it can provide rewards to the agent based on the agent’s actions. During the test, the agents interact with a dynamic environment to maximize the predefined reward. In particular, the agent can send an action signal to the environment and receive a reward corresponded to the current action from the environment. Such test can be regarded as an interactive process between agent and environment and certain information can be collected by observing the process.

3. Quality–time–complexity intelligence measuring model

In this section, we first analyze the correlations between the reward quality and the two other factors. And next, we will introduce our QTC intelligent measure model in detail.

3.1 Major factors for measuring intelligence

By observing the interaction process between the agent and the environment, we abstract three major factors that determine the performance of agent during the intelligent test as follows:

1. **Reward** is a sequence of the reward, which is derived based on the actions taken by the agent, and it is quantified by calculating the Expected Accumulated Reward (EAR) of the reward sequence.

2. **Time** is the timestamp of the rewards, which can represent the timeliness of the agent’s actions.

3. **Environment** is the complexity of the test environment, which can be computed and these environments can adjust based on evaluating the agents’ actions.

To evaluate the correlation among the three factors, we conducted two experiments with seven agents involved. In the first experiment, we performed the same intelligent test on four agents. During the tests, we observe that the variation of EAR was obtained by the tested agent by progressively increasing the complexity of environment. The results of the first experiment (Figure 2) indicate that although the EARs of the four agents change with different patterns when the complexity of environment increases, all of them converge when the complexity of environment is above 21. In the second experiment, we performed the same intelligent test on three agents. The third agent in this experiment configured invokes random actions. The result (Figure 3) shows that the three EARs all increase with time. Moreover, the EARs of the first two agents converge as numbers of interactions increase. Based on the results of the two experiments, we model the correlations of the three factors as a diagram shown in Figure 4. As shown in the figure, time and complexity of environment both correlate with the Reward factor, as the EAR converges when time and complexity of environment increase to a certain threshold.

![Figure 1. The interaction between the agent and the environment](image)
After analyzing the relationship between them, our next question is how to calculate the EAR.

3.2 The reward for each interaction
As the goal of intelligent measure is to calculate the value of the reward, our first task is to calculate the reward for each interaction. According to the intelligent test designed by Legg and Hutter (2006), a complete interaction between the agent and the environment includes two steps: the agent sends an action to the environment and the environment evaluates the action and returns a reward to the agent. For instance, in the Turing test, a complete interaction includes a question asked by the agent and an answer responded by a human.

In an intelligent test where the finite number of interactions occurs within a finite time period, we define the reward $R_i$ of interaction $i$ as:
\[ R_i(t) = \left(1 + \frac{1}{mt}\right)^t \]  

(1)

where \( m \) is the complexity of environment and \( t \) is the time when the action \( i \) is invoked. This equation is designed according to the trend shown in Figures 2 and 3.

When duration of an intelligent test is infinite long so that \( t \to \infty \), the limit value of \( R_i \) is:

\[ \lim_{t \to \infty} R_i = \lim_{t \to \infty} \left(1 + \frac{1}{mt}\right)^t = e^m \]  

(2)

The result in equation (1) is a constant when conforms the convergences shown in Figure 2. Hence, such result proves the correctness of equation (1).

The complexity of environment \( m \) can be calculated according to the algorithm information theory by using Levins Kt complexity (Li et al., 2008; Levin, 1973) as follows:

\[ m(p, \pi) = \min\{l(p) + \log \text{time}(\pi, p)\} \]  

(3)

where \( p \) represents the action and \( \pi \) represents the agent.

By substituting equation (3) into equation (1), the reward \( R_i \) of interaction \( i \) can be calculated as:

\[ R_i(t) = \left(1 + \frac{1}{m_{pi} t}\right)^t \]  

(4)

3.3 Intelligence measuring model

In this paper, we measure the intelligence of agent \( \pi \) by calculating the EAR obtained by agent \( \pi \) within a predefined period \( t \). Hence, the objective of intelligence measure model is to accurately calculate the EAR. According to the intelligence measuring model introduced by Hernndez-Orallo and David (2010), the calculation of EAR can be based on the sum of the average rewards obtained by agent \( \pi \) within a predefined period \( t \) (defined as \( V_{\mu}^\pi \)). The equation to calculate \( V_{\mu}^\pi \) is as follows:

\[ V_{\mu}^\pi := E \left( \sum_{i=1}^{n} R_i \right) = \frac{1}{n_i} \sum_{i=1}^{n_i} \left(1 + \frac{1}{m_{pi} t}\right)^t \]  

(5)

where \( n_i \) is the total number of interactions and \( \mu \) is the identity of environment. Based on equation (5), we derived the EAR from Legg and Joel’s study (2013):

\[ \Upsilon := \sum_{\mu \in E} 2^{-k(\mu)} V_{\mu}^\pi \]  

(6)

By substituting equation (5) into equation (6), we can obtain the value of EAR as:
where the environment $\mu$ belongs to the environment set $E$ which includes all computable reward bounded environments, and $K(\cdot)$ is the Kolmogorov complexity.

By combining equations (1)-(7), we can eventually propose our intelligence measure model as:

$$\Upsilon(t; \theta) := \frac{1}{n_i} \sum_{\mu \in E} 2^{-k(\mu)} \sum_{i=1}^{n_i} \left( 1 + \frac{1}{m_{pi} t} \right)^t$$

(8)

subject to $t > t_0$

(9)

where $\theta = (\mu, \pi)^T$ is the parameter of EAR and $\mu$ is the identity of environment.

4. The results analysis of the model
In this section, we implement our proposed intelligent model and then conduct an experiment to evaluate the performance of our model.

4.1 The algorithm of quality–time–complexity universal intelligence measurement
The implementation of our proposed model is described by the following pseudocode:

**Algorithm 1 Universal intelligence test**

**Input:** $t$ (the time of the interaction), $p$ (interactive behavior)

**Output:** a real number (the rewards of the interaction between the agent and the environment)

1: Calculate complexity of environment $m$ based on (2).

2: Calculate the reward for action $R_i$ based on (3).

3: Calculate the expected sum of the rewards $V^\pi_\mu$ based on (4).

4: Calculate the Expected accumulated reward $\Upsilon$ based on (6).

5: **return** $\Upsilon(t; \theta)$

Based on the pseudocode introduced above, we performed a simulation to visualize the correlation between the Expected Cumulated Reward, time and the complexity of environment.

According to the simulation result shown in Figure 5, it can be seen that the Expected Cumulated Reward increases with time, and decreases significantly as the complexity of the environment increases.

4.2 Experimental analysis
In this section, an example of implementing the proposed intelligent measure is introduced in detail. Consider a test setting where a chimpanzee (the agent) can press one of the three
buttons ($A = \{B1, B2, B3\}$). Rewards can be giving the agent either a banana or nothing ($R = \{0, 1\}$). The observation set is derived from an environment where a ball must be put into one of the three cells ($O = \{C1, C2, C3\}$). We start the test by giving a banana to the chimpanzee, which indicates that the first reward is 1. The observations are randomly generated with a uniform distribution with respect to $O$ so that the rewards are determined accordingly. The behavior patterns of the agents are designed as follows.

The first chimpanzee $\pi_1$ is much more likely to press button $B1$, i.e. $\pi_1(B1|X)$ for all sequences $X$. Consequently, the performance of $\pi_1$ in this test is:

$$E(V^{\pi_1}) = E_{n_i \to \infty} \left( \frac{\sum_{k=1}^{n_i} R_k^{\mu, \pi_1}}{n_i} \right) = \frac{2}{4} \lim_{n_i \to \infty} \frac{n_i}{n_i} + \frac{2}{4} \lim_{n_i \to \infty} \frac{0}{n_i} = \frac{1}{2}$$  \hspace{1cm} (11)

The second chimpanzee ($\pi_2$) behaves randomly. Hence, the performance of $\pi_2$ is:

$$E(V^{\pi_2}) = E_{n_i \to \infty} \left( \frac{\sum_{k=1}^{n_i} R_k^{\mu, \pi_2}}{n_i} \right) = \frac{3}{3} \left( \frac{2}{4} \lim_{n_i \to \infty} \frac{n_i}{n_i} + \frac{1}{4} \lim_{n_i \to \infty} \frac{-n_i}{n_i} + \frac{1}{4} \lim_{n_i \to \infty} \frac{-n_i}{n_i} \right) = 0$$  \hspace{1cm} (12)

By comparing the performance between the two agents, we can conclude that agent $\pi_1$ is better than agent $\pi_2$ during this test.

### 5. Conclusion and future work

Traditional human intelligence and machine intelligence are difficult to be described by the form of intelligence in the current environment, and have great limitations. In this paper, we propose a universal intelligent measure approach: QTC intelligence measure approach. We abstract three major factors for intelligence measure as quality, time and complexity of environment. Correlation of the three factors is estimated by conducting two experiments so

![Figure 5](image.png)

**Figure 5.** Relationship between the three major factors of intelligent measurement, where $x$-axis represents the complexity of environment, $y$-axis represents the time and $z$-axis represents the EAR

**Note:** EAR increases with time and decreases with the complexity of the environment
that the intelligence measure mode can be designed accordingly. Based on the intelligent measure model, we can quantify the intelligence of an agent by calculating the EAR achieved by the agent during an intelligent test. In future, we plan to design and implement a set of comprehensive experiments to evaluate the performance of our measurement model.

References


Further reading


About the authors
Jing Liu is currently pursuing MS at the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China. Her current research interests include crowd science, intelligence measure, machine learning and pervasive computing.

Zhiwen Pan received BS degree from the Purdue University Calumet, in 2012, and the MS and PhD degrees from the University of Arizona, in 2014 and 2017, respectively. He is currently an Assistant Research Fellow in Research Center for Ubiquitous Computing System, Institute of Computing Technology, Chinese Academy of Science. His current research focuses on cybersecurity, secure critical infrastructures, anomaly detection and context awareness.
Jingce Xu was born in 1994. He received BSc in Software Engineering from Nankai University, China, in 2016. He is currently working toward MSc in Computer Application Technology at the Institute of Computing Technology Chinese Academy of Sciences. His research interests include multimedia communication and networking, video transmission and video QoE optimization.

Bing Liang is currently pursuing PhD at the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China. His current research interests include multimedia communication and networking, video transmission, game theory and pervasive computing.

Yiqiang Chen received BSc and MS degrees from the University of Xiangtan, Xiangtan, China, in 1996 and 1999, respectively, and PhD degree from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2002. In 2004, he was a Visiting Scholar Researcher with the Department of Computer Science, Hong Kong University of Science and Technology (HKUST), Hong Kong. He is currently a Professor and Director with the Pervasive Computing Research Center, ICT, CAS. His research interests include artificial intelligence, pervasive computing and human–computer interface.

Wen Ji received MS and PhD degrees in communication and information systems from Northwestern Polytechnical University, China, in 2003 and 2006, respectively. She is a Professor in the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China. From 2014 to 2015, she was a Visiting Scholar at the Department of Electrical Engineering, Princeton University, USA. Her research areas include video communication and networking, video coding, channel coding, information theory, optimization, network economics and ubiquitous computing. Wen Ji is the corresponding author and can be contacted at: jiwen@ict.ac.cn
Abstract

Purpose – Participatory wireless sensor networks (PWSN) is an emerging paradigm that leverages existing sensing and communication infrastructures for the sensing task. Various environmental phenomenon – P monitoring applications dealing with noise pollution, road traffic, requiring spatio-temporal data samples of P (to capture its variations and its profile construction) in the region of interest – can be enabled using PWSN. Because of irregular distribution and uncontrollable mobility of people (with mobile phones), and their willingness to participate, complete spatio-temporal (CST) coverage of P may not be ensured. Therefore, unobserved data values must be estimated for CST profile construction of P and presented in this paper.

Design/methodology/approach – In this paper, the estimation of these missing data samples both in spatial and temporal dimension is being discussed, and the paper shows that non-parametric technique – Kernel Regression – provides better estimation compared to parametric regression techniques in PWSN context for spatial estimation. Furthermore, the preliminary results for estimation in temporal dimension have been provided. The deterministic and stochastic approaches toward estimation in the context of PWSN have also been discussed.

Findings – For the task of spatial profile reconstruction, it is shown that non-parametric estimation technique (kernel regression) gives a better estimation of the unobserved data points. In case of temporal estimation, few preliminary techniques have been studied and have shown that further investigations are required to find out best estimation technique(s) which may approximate the missing observations (temporally) with considerably less error.

Originality/value – This study addresses the environmental informatics issues related to deterministic and stochastic approaches using PWSN.

Keywords Regression, Approximation, Environmental informatics, Spatial-Temporal profile

Paper type Research paper

1. Introduction

Participatory wireless sensor networks (PWSN) (Bruke et al., 2006; Miluzzo et al., 2010) is an emerging paradigm where instead of deploying dedicated infrastructures, existing
sensing (handheld devices) and communication infrastructures (i.e. Cellular, Wifi, WiMax, internet) are used to perform the sensing task. The ubiquity of mobile phone handsets is the key driver of the above mentioned paradigm. In addition to video-image (camera) and acoustic (microphone) sensors, one may attach other sensors (using Bluetooth) to the handset. According to Kansal et al. (2007) can enable various applications including noise (and, or) air pollution monitoring, item price sharing, locating urban area parking lots, etc.

First community noise mapping application is considered. Noise pollution is a serious concern in urban areas. Noise maps of some cities (e.g. the UK) available on the Web are generated using computer-based modeling techniques; therefore, they do not precisely represent the actual noise level. Santini et al. (2008) have proposed noise pollution monitoring using dedicated sensors, but deploying a large network of dedicated nodes in a city may not be possible. Alternatively, a highly scalable way is to use mobile phones as acoustic sensors when exposed to open air (Kanjo, 2010). The noise-level samples acquired can be relayed to a central data sink to build so-called spatio-temporal noise profile that can be mapped on geographic (Google) map to visualize noise pollution of a given community. Various noise level indicators are quarried from the built profile including $L_{eq,T}$ (equivalent noise level for time period $T$), $L_{i}$ (noise level at time interval $i$), $L_{ips}$, $L_{per cent,1h}$, $L_{per cent,day}$, $L_{day}$, $L_{evening}$. The profile may serve many purposes such as selecting quite place to live in the city, trend analysis of noise pollution, urban planning, help in selecting location (for a new entertainment center, pub, industry, hospital or residential block), or to decide whether to hire acoustic engineers for the survey of the chosen location. The basic design principle of PWSN is the user’s contribution (i.e. people carrying mobile devices and willing to participate for sensing task). Furthermore, irregular geographic distribution, unpredictable mobility of people and lack of synchronization of acquired data samples result in loss of coverage. The irregular spatio-temporal sampling and nature of $P$ make the problem significantly different from the missing data problems in other data acquisition systems. Mobile phones are not meant to be used for dedicated sensing, thus negotiating sampling rates (and intervals) with heterogeneous set of users may not work here. Therefore, to construct the complete spatio-temporal (CST) profile of the phenomenon $P$, estimation of missing data samples is required such that the data remain valuable in spatio-temporal context.

The study aims to investigate the noise data in environmental informatics which is collected while logging from various sensing devices. It has been observed that there is a possibility of some missing data or some incoherent noise intervention at the time of collecting environmental data from different types of logging devices. Attempt has been made to minimize the noise and measure the appropriate missing values. PWSN have been used for measurements and large amount of homogeneous data have been retrieved during the logging process. The study proposes spatio-temporal and geo-statistics mechanisms of data analysis for the real-time cleaning, filtering and mapping. Environmental data informatics is an emerging discipline which helps to clean environmental data. This study addresses the environmental informatics issues related to deterministic and stochastic approaches using PWSN. The focus is on the comparative study of parametric and non-parametric regression techniques to estimate the missing data samples.

The main focus is on the comparative study of parametric and non-parametric regression techniques to estimate the missing data samples. The simulation results show that the non-parametric approach outperforms the parametric techniques thus suits well for estimation task in PWSN. Noise pollution monitoring is used as an illustrative example; however, same estimation techniques are applicable to other applications such as road traffic monitoring, air pollution monitoring and CO2 monitoring.
The rest of the paper is organized as follows. Section 2 gives an overview of the related work; Section 3 discusses the noise profile construction problem and proposes a solution with a brief introduction to parametric and non-parametric techniques, i.e. deterministic and stochastic approaches along with introduction to the temporal regression techniques. Section 4 describes a prototype implementation of smart phone handset as acoustic sensor. Section 5 presents the simulation model and estimation results. Finally, Section 6 details the findings of the study and provides the future research directions.

2. Related work
The main focus of the current research in WSNs is to use dedicated infrastructures (Akyildiz et al., 2002). An alternative paradigm for WSNs, i.e. PWSN has emerged recently. A few recently proposed projects include Participatory Sensing (Bruke et al., 2006), SenseMart–Sensing Data Market (Chou et al., 2007), Mobile GeoSensing (Kanjo, 2010), MetroTrack (Ahn et al., 2010), SoundSense (Lu et al., 2009) and Bubble-Sensing (Lu et al., 2010). PWSN has also been introduced by some other groups of people including Schweizer et al. (2011, 2012), Maisonneuve et al. (2010) and Baykasoglu et al. (2016) for participatory noise map generation.

Missing (or unobserved) data refers to the difference between data planned to be collected and data collected in actual. This problem has been widely studied in different fields such as statistics, databases, field of medicine (Lakshminarayan et al., 1999; Longford, 2005; Little and Rubin, 2014). Some of the literature about data sample estimation in sensor networks is briefly discussed as under.

There have been number of attempts to deal with missing data in WSN. Jiang and Gruenwald (2007) and Nag et al. (2015) propose data estimation technique using association rule mining in the data streams. The problem with these techniques is that the missing behavior in PWSN is mainly because of non-availability of data producing nodes; therefore, association rule mining (Indira and Kanmani, 2015) approaches may not always work which is suitable for data streams with comparatively small-scale missing data. Ensemble techniques have also been used for handling missing data (Mohammed et al., 2006). The ensemble approaches work well when data are corrupt and use weighted majority approach for classification but are not able to handle situations with a large number of missing features and hence work only when limited number of sensors malfunction.

Elnahrawy and Nath (2004) have proposed an online learning of spatio-temporal correlations and utilized them to discover outliers, approximate missing values and detect faulty sensors, however, the assumption behind their work is dense network with redundant data and correlated readings for coverage and connectivity.

Jiang and Gruenwald (2007) and Sutha and Dhanaseelan (2017) proposed the use of association rule mining for estimation in data streams. However, missing data behavior in PWSN (because of non-availability of users), association rule mining approaches (suitable for small-scale missing data) may not always work. The authors proposed spatio-temporal correlations learning to discover outliers, approximate missing values and detect faulty sensors assume a dense network with redundant readings for coverage and connectivity. To build Worldwide Sensor Web, Balazinska et al. (2007) have raised various data management issues including missing data and used interpolation to fill them. Guestrin et al. (2004) have proposed parametric regression to model sensor data. Deshpande and Madden (2006) have also used the aforementioned approach to estimate missing data values. Although parametric model may work well, however, the approach may not capture unforeseen situations in dynamic environments (Christin et al., 2011). Therefore, fixing a model may result in loss of flexibility to explore, analyze, and decide based on available real-time data.
Therefore, a flexible, data dependent (nonparametric) method to estimate missing samples exploiting their relationship with the sensed data has been proposed (Xu et al., 2016).

3. Sensor coverage and spatio-temporal profile reconstruction

The environmental noise is highly dynamic and fluctuates depending upon nature (frequencies and loudness) of the sound sources (Bies and Hansen, 2009). Typical sound sources include transportation systems (road, rail and air traffic), construction work, factory and audio entertainment systems. To capture most of the noise variations, finely grained sampling is required. Therefore, to acquire noise samples at grid points, entire region of interest can be divided into a regular grid with certain dimensions (say 5 × 5 m). Similarly, the temporal granularity can also be fixed at regular intervals to build CST noise profile of the region.

As in PWSN, granular sampling from the whole region may not be possible; therefore, a mechanism is required to estimate the unobserved data points. The focus is to construct spatial profile for unit time $t_i$. Once constructed, the profile may be used to estimate unobserved data points at the next time stamp $t_j$ along with the observations collected from the region at time $t_i$. The preliminary results are provided for the estimation task in time dimension with the point of view that historical data sets may help to reduce the estimation error while approximating an unobserved data point.

The attributes ($A$) in each data sample $S_i$, i.e. geographic location $x \in \mathbb{R}^2$ (or $x \in \mathbb{R}^3$ for 3D case); $t$, time interval; equivalent noise level, $L_{eq}$, during $t$; and time stamp can be represented as:

$$S_i \rightarrow A_j = (a_1 \ a_2 \ \ldots \ a_m).$$

Received data samples at an instance $I$ can be represented as:

$$I = \begin{pmatrix}
a_{11} & a_{12} & \ldots & a_{1m} \\
a_{21} & a_{22} & \ldots & a_{2m} \\
\vdots & \vdots & \ddots & \vdots \\
a_{n1} & a_{n2} & \ldots & a_{nm}
\end{pmatrix}$$

Assume expected number of samples, $N_{exp}$ required for the profile construction of $P$ (based on spatial granularity requirement of $P$) is known, and $N_{recv}$ be the number of samples received, missing data samples problem can be expressed as: $N_{recv} < N_{exp}$. Thus, completeness of the spatial profile (matrix $I$) can be determined using:

$$r = \frac{N_{recv}}{N_{exp}}$$

Where $r \in \mathbb{R} \land 0 \leq r \leq 1$ assuming $N_{recv} \leq N_{exp}$ (always); hence, the ratio $r$ depicts quantitative measure of missing samples as shown in Figure 1.

The objective of the estimation task is to approximate the missing entries in matrix $I$ using regression techniques as mentioned in Algorithm 1 (Figures 2 and 3).

3.1 Community noise map and estimation techniques

Referring to the noise map construction and noise (sound) propagation in an ideal case follows sound propagation obeys Inverse Square Law Bies and Hansen (2009). The law
states that in a "free field", sound will fall at the rate equal to the inverse of square of the distance (from the source). In fact, there are many factors which affect the spherical propagation of sound. As a result, the sound signal attenuates is not allowed to dissipate as Inverse Square Law suggests. The outdoor sound propagation depends on several factors which include refraction (because of gradients of wind and temperature), reflection (at ground, building, forests etc), diffraction (at obstacles e.g. buildings, hills, big signboards/ screens etc), scattering (because of atmospheric turbulence), absorption because of humidity. The equation for sound signal propagation is given as:

\[ L_p = L_w - 20 \log r - 11 \text{ (dB)} \]  

where \( L_p \) is the sound pressure level, \( L_w \) is sound power level and \( r \) is the distance from the source in meters. After taking into account the above-mentioned factors affecting the sound propagation, equation (3) becomes:

\[ L_p = L_w - 20 \log r - 11 - A_{abs} - A_E \text{ (dB)} \]  

where: \( A_{abs} \) is atmospheric absorption and \( A_E \) is excess attenuation (dB). The total attenuation \( A_E \) (dB) is a combination of all effects:

\[ A_E = A_{\text{weather}} + A_{\text{ground}} + A_{\text{turbulence}} + A_{\text{barrier}} + A_{\text{vegetation}} + \ldots \]  

**Figure 1.** Data samples received effecting profile completeness

**Figure 2.** Algorithm: missing data samples(s) estimation
These terms are quantified using different set of equations, which can be studied in detail Dunn et al. (2015). From the above-mentioned sound propagation model, it is obvious that accurate measurement of all the factors involved in the sound propagation is very difficult, if not impossible, to calculate noise level at certain location from the source as to exactly measure all the factors involved.

3.2 Estimation in spatial dimension
The inability to acquire maximum (desired) coverage in the data collection (of noise level samples) leads to insufficient and missing data problem as the valuable data required to build the spatio-temporal profile of the whole region may not be available at required resolution. As discussed earlier, to maximize the coverage from the data collected, estimation mechanism is required to estimate the noise level at the locations where data have not been collected. For the significance of the estimation, the approximation procedure should take into account the data collected from other locations. Given that the noise level data are missed at certain location, in fact, there are two ways to estimate the noise level: deterministic and stochastic.

**Figure 3.** Regression analysis: parametric vs non-parametric (Kernel) regression

Notes: (a) 10% unobserved data points; (b) 30% unobserved data points; (c) 60% unobserved data points; (d) 70% unobserved data points
For the deterministic solution, all the parameters should be known which have been discussed above along with the sound pressure level of the sound source(s). The other factors critical for deterministic solution are types (point, cylindrical or plane) and nature (uni or omni directional) of the sound source(s) responsible for propagating noise in the region. By conducting experimental studies, it can be concluded that from given set of noise level data from users, it is complicated procedure to determine the noise level of the sound source(s), and it is hard to determine the number of sound sources involved. Even if it is known, there are many other factors like type and nature of sound source(s), factors involved in sound propagation, etc. that need to be determined to calculate noise level at a given location, which is hard to get in real-world scenarios. Therefore, to determine noise level at a given location using deterministic approach is hard. This leads the study to more realistic approach to estimate the missing noise level data at a given location, also known as stochastic (non-parametric) approach. According to this approach, the noise level at a given location can be estimated based on the data gathered from the neighboring locations. Regression analysis is applied using both parametric and non-parametric techniques for estimation task. The techniques are briefly discussed below.

3.2.1 Parametric regression. In parametric regression, the aim is to find the best-fit equation for the data, e.g. in case of Linear regression, the relationship between dependent (response, unknown) and independent (explanatory, received data) variables are assumed to be linearly defined:

\[ y_i = ax_i + b + \epsilon_i, \quad i = 1, \ldots, n \]  

where \( y_i \) is the response, \( x_i \) is the explanatory variable, \( \epsilon_i \) is random error, \( a \) is slope and \( b \) is the y-intercept of the regression line George and Collins (2003). Equation (6) describes a line in \( x-y \) plane with \( n \) pairs \( (x_i, y_i) \) of data samples on that plane.

The response variable may depend on a nonlinear function of explanatory variables (e.g. realistic carbon emission predicted future global warming is not expected to be a simple linear function of time). We used Polynomial regression to estimate missing data values (George and Collins, 2003).

\[ y = f(x) + \epsilon \]  

where \( f(x) \) is the polynomial (aka basis) function and can be of order 1, 2, 3, \ldots \( n \). We have also used Cubic spline interpolation; details can be seen in Deboor’s (1978) study.

3.2.2 Non-parametric regression. In contrast, non-parametric methods can be used to build up an overall model of the data based on simple and intuitive local models, which can be used for the scenarios with unknown (or difficult to fit) theoretical models (Hardle, 1990). The aforementioned approach suits to the PWSN estimation problem because:

- It allows to tailor estimation problem to the local characteristics of the observed data.
- It is more flexible (and robust) in setting up model parameters dynamically and does not require the data to follow equally spaced sampling.

Kernel regression is used for non-parametric estimation of unobserved data samples, as the method is simple to analyze and is mathematically well understood (Hardle, 1990). The key idea is to use weighted average of the observed data values, where at a given location \( x \in \mathbb{R}^2 \) (or \( x \in \mathbb{R}^3 \)), noise level at \( x \) is conditionally independent of far away locations, defined as:

\[ \hat{m}(x) = \frac{1}{n} \sum_{i=1}^{n} W_i(x) v_i \]  

where

- \( W_i(x) \) are weights that depend on the distance between \( x \) and \( x_i \) (can be chosen as Gaussian or Epanechnikov).
where $\hat{m}(x)$ is the estimated noise level at $x$, \( \{W_i(x)\}_{i=1}^{n} \) denotes sequence of weights that may depend on all the observations at \( \{x_i\}_{i=1}^{n} \) and $v_i$ is the noise level observed at $x_i$. For $W$, the weighting procedure as proposed by Nadaraya–Watson (Nadaraya, 1964), was applied and thus Nadaraya–Watson Kernel Regression (NWKR) is given as:

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K_h(x - x_i)v_i}{\sum_{i=1}^{n} K_h(x - x_i)} \quad (9)$$

where $K_h = K(\frac{x-x_i}{h})$ is the kernel function and $h$ is the neighborhood bandwidth of $x$ used to assign weights to the neighboring data samples. Hardle (1990) has shown that $h$ effects the estimation accuracy, therefore, needs to be carefully selected. For Kernel function $K$, we defined it to be Gaussian, as the choice of $K$ has less impact on estimation accuracy (Hardle, 1990). Gaussian Kernel is defined as:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{(-u^2/2)}, \text{where } u = (x - x_i)/h \quad (10)$$

3.3 Estimation in temporal dimension

In this section, the temporal profile reconstruction has been discussed. The data samples acquired can be represented as time-series since the data samples are gathered at successive time stamps for a time period. It is worth mentioning that the static analysis of the acquired data samples for a given time period, i.e. the techniques are applied to the data samples acquired once sampling for the entire time period has finished. For the estimation of missing data values, we have used the popular time-series prediction models such as moving average and autoregressive (AR) models. The purpose of the study is to determine the accuracy of the time-series prediction models and their applicability in the context of participatory sensing, where data samples from a given region may not be synchronized with each other spatially and temporally.

3.3.1 Moving average. Moving average is most commonly used smoothing techniques, which is often used to provide time-series prediction. The techniques work by reducing the randomness of the (time-series) data samples, and may lead to better prediction. Moving average model of order $q$ can be given as under:

$$X_t = \mu + \epsilon_t + \sum_{i=1}^{q} \Theta_i \epsilon_{t-i} \quad (11)$$

where $\Theta_1, \ldots, \Theta_q$ are the parameters of the model, $\mu$ is the expectation of $X_t$ and $\epsilon_t, \epsilon_{t-1}, \ldots$, are the noise error terms.

As in general, time series analysis, simple moving average models performed better than higher window size models. The question arises that, do other advanced time-series prediction models perform any better than moving average models for our estimation problem. To investigate this issue, we evaluate the performance of advanced time-series prediction models. In particular, we investigate the performance of the AR and AR moving average (ARMA) models.

3.3.2 Autoregressive model. The AR model considers a value at time $t$ based on the linear combination of prior values (i.e. forward prediction) and upon the combination of subsequent
values (i.e. backward prediction). In other words, the AR model uses the acquired data values at a particular grid location to predict the values at next time interval. Equation (12) refers to mathematical representation of the AR model of order \( p \).

\[
X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_i
\]  

(12)

where \( \phi_1, \ldots, \phi_p \) are parameters, \( c \) is constant and the random variable \( \epsilon_i \) is the noise.

3.3.3 Autoregressive moving average model. Using both moving average and AR models, ARMA model may also be used for the estimation task. In other words, the model contains both \( AR(p) \) and \( MA(q) \) models. The mathematical notation of the ARMA model is given in equation (13).

\[
X_t = c + \epsilon_t + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{i=1}^{q} \Theta_i \epsilon_{t-i}
\]  

(13)

Where, \( c \) is constant, and the variable \( \epsilon_i \) is the noise.

4. Acoustic sensor implementation
To sense the noise level at a given location a program in Microsoft Visual C# was developed for hand held smart phones – PDA’s to find noise level (dB) at a location for the sound signals captured using built-in microphones. The implementation involves calculation of noise exposure level (\( L_E \), a metric to measure noise level, with reference time interval set to 1 second) as given below:

\[
L_E = 10 \log_{10} \left\{ \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} \frac{p^2(t)}{P_0^2} dt \right\}
\]  

(14)

As discussed above, the implementation involves calculation of the noise exposure level \( L_E \) as per equation (14), where reference time interval can be set up to the desired number. Currently, the reference time interval has been set to 1 s. The implementation involves various steps, which are briefly stated as under.

By using a set of library functions, which give access up to microphone level for PDA’s (Mitchell, 2007), the sound samples were acquired to determine noise level (dB). As microphone converts the sound signals into electrical signals, the energy level of the captured sound samples was calculated by applying fast Fourier transform, followed by a weighting filter to assign weights to the transformed signals and then Parseval’s relation (Mitra and Kuo, 2006), which is used to estimate the signal energy as shown in equation (15).

\[
E = \sum_{n=0}^{N-1} (x[n])^2
\]  

(15)

The final dB(A) is calculated using the relationship:

\[
\text{Signal Level in dB(A)} = 10 \log_{10} \left( \frac{E}{E_{ref}} \right)
\]  

(16)
where $E_{ref}$ is the (constant) reference signal level, equation (16) becomes:

$$\text{Signal Level in dB(A)} = 10\log_{10}E + C$$

(17)

where $C$ is calibration constant and can be determined by acquiring samples and cross reference with the sound level meter readings.

5. Simulation model and results
MATLAB was used to simulate noise profile construction scenario from the sparse sampling to study the accuracy of various regression techniques. Vehicular traffic noise propagation (Makarewicz, 1998) model was considered to disperse the noise from the center of the main road carrying metropolitan traffic (over 1500 vehicles per hour in the peak period). Vehicles on the road follow Poisson arrival rate (Songchitruksa and Hard, 2008). The sensing nodes (people) are positioned at grid of 1 m size on the sidewalk along the road (of dimensions 100 × 4 m). The ground truth (noise level) was gathered at all the grid points. Moreover, different sparse data sets were generated from the ground truth by randomly omitting the samples from grid points at certain (increasing) percentage e.g. 10, 20 and 30 per cent.

It is worth noting that realistic model has been used for environmental noise propagation, as experimental settings on even small-scale require quite a few resources and the results might get affected by environmental and or human influences. However, in the next step the empirical study is supposed to be done to compare our findings with simulation results.

5.1 Profile reconstruction in spatial dimension
This section discusses the simulation results for the estimation in the spatial dimension. The regression techniques were implemented, and applied on the generated sparse data sets to estimate the omitted (unobserved) traffic noise data samples.

To cross-validate (with the ground truth) and measure the accuracy of the estimation results, the standard metric mean absolute percentage error (MAPE) (Hardle, 1990) was used for the 100 runs (of each technique) with randomly omitted samples. Figure 6 shows the comparison results of various regression techniques i.e. Linear Regression (LR), polynomial regression of degree 2 to 5 (PR2, ..., PR5), Cubic Spline (CS) and Gaussian Kernel regression (with 0.25, 0.4 and 0.5 meters of $h$). The data sets used for the comparison contain 10, 30, 60, and 70 per cent unobserved points respectively. The results shows that the kernel regression outperforms the parametric regression techniques, because of the fact that modeling $P$ with real-time unpredictable behavior with fixed parametric models is hard, therefore, non-parametric models should be preferred.

In further simulations, it was observed that MAPE is inversely proportional to $h$ [shown in Figure 4(a)] because, for the higher value of $h$, more data points are involved for the estimation and the weights of the observed data points are smoothed decreasing the overall MAPE. It can be seen that increase in MAPE up to a certain percentage of unobserved data points is not very significant because grid points are closely located and noise variation is less because of single type of sound source being used. Figure 4(b) presents 90 per cent confidence interval of MAPE that increases with the increase in the percentage of unobserved data points.

5.2 Profile reconstruction in time dimension
The experimental results for the temporal profile reconstruction using the techniques discussed above 3.3 are discussed here. Upon analyzing the time series data for the grid points for the current simulation scenario, results for increasing percentage of unobserved
Data values suggest that moving average has failed to model the series, as indicated by the accuracy metric, i.e. mean percentage error (MPE) aka mean prediction error. The results can be generalized as shown in the Figure 5 for this particular simulation scenario, as the MPE for moving average converges for the data samples, where each data sample location is the grid point location in the simulations.

Figure 6(a) plots the mean prediction error as a function of percentage of missing data samples, for different moving average window sizes. The increase of missing data samples results in a linear increase of the MPE. For low percentage, missing data the higher window sizes of moving average have the ability to considerably reduce the MPE, whereas, at the higher percentage of missing data, the reduction of MPE is less. To better describe the effect of increasing the window size of the moving average at higher missing data samples, Figure 6(b) presents the MPE as a function of moving average window sizes. It is evident that for lower percentage of missing data samples, the simple moving average models...
(e.g. moving average with window size of 5 and 10) reduce the MPE of prediction. However, higher window size models have no significance in reducing the average percentage error.

Figure 7(a) plots the mean prediction error as a function of missing data samples, for different orders of AR model. As the case with moving average models, increasing the percentage of missing data samples results in a linear increase of the mean prediction error. However, higher-order AR models increase the MPE. The results are found surprising, as it suggests that the higher-order AR models are not fitting the data under investigation, which may point out to the fact that the correlation between the different points of the time-series does not carry any significance. To conclude this, further investigations are required and will be done as part of the future work. Figure 7(b) is plotted to further investigate the effect of increasing the order of AR model on mean prediction error, the figure confirms that the higher-order AR models increases the MPE.

Although the synthetic data generated by the vehicular propagation model is expected to be noise-free, yet the performance of ARMA models is being investigated. ARMA models usually applied in the cases where data are affected by noise, as it enhances the prediction of the AR model by incorporating the noise in the estimation process. Figure 8(a) and (b)
clearly confirms that the results of ARMA are very similar to the results gathered by the AR model, which emphasizes that the data is noise-free.

6. Conclusion and future work
The estimation of unobserved data samples for PWSN applications, especially environmental noise pollution monitoring, in particular, has been discussed. Acoustic sensor implementation and practical utilization of microphones as acoustic sensor for the smart devices has also been discussed. Various estimation techniques for both spatial and temporal profiles have been used to approximate missing observations, and the results for the techniques have been compared. The basic model for expected, received and missing samples provided. For the task of spatial profile reconstruction, it is shown that non-parametric estimation technique (kernel regression) gives a better estimation of the unobserved data points. In case of temporal estimation, few preliminary techniques have been studied and have shown that further investigations are required to find out best estimation technique(s), which may approximate the missing observations (temporally) with considerably less error. This task will be done as part of any future study. Furthermore, in the future study the following will be investigated:

- Integration of spatial and temporal profile estimates to produce CST (urban noise) profile in a region, as both spatial and temporal observations seem to correlate in other words, may assist each other for better approximations.
- The approximation algorithm developed here will be implemented and further improved on more realistic scenarios, (as in reality it is hard to fix the location of mobile users at the grid points) to find out how the proposed scheme works in a realistic environment.

References


**Corresponding author**
Arshad Muhammad can be contacted at: amuhammad@soharuni.edu.om

For instructions on how to order reprints of this article, please visit our website: www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: permissions@emeraldinsight.com
Towards data-driven software engineering skills assessment

Jun Lin
The Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY), Nanyang Technological University, Singapore, and Alibaba-NTU Singapore Joint Research Institute, Singapore

Han Yu
School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore

Zhengxiang Pan
Interdisciplinary Graduate School, Nanyang Technological University (NTU), Singapore

Zhiqi Shen
School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore, and

Lizhen Cui
School of Computer Science and Technology, Shandong University, China

Abstract

Purpose – Today’s software engineers often work in teams to develop complex software systems. Therefore, successful software engineering in practice require team members to possess not only sound programming skills such as analysis, design, coding and testing but also soft skills such as communication, collaboration and self-management. However, existing examination-based assessments are often inadequate for quantifying students’ soft skill development. The purpose of this paper is to explore alternative ways for assessing software engineering students’ skills through a data-driven approach.

Design/methodology/approach – In this paper, the exploratory data analysis approach is adopted. Leveraging the proposed online agile project management tool – Human-centred Agile Software Engineering (HASE), a study was conducted involving 21 Scrum teams consisting of over 100 undergraduate software engineering students in multi-week coursework projects in 2014.

Findings – During this study, students performed close to 170,000 software engineering activities logged by HASE. By analysing the collected activity trajectory data set, the authors demonstrate the potential for this new research direction to enable software engineering educators to have a quantifiable way of understanding...
their students’ skill development, and take a proactive approach in helping them improve their programming and soft skills.

Originality/value – To the best of the authors’ knowledge, there has yet to be published previous studies using software engineering activity data to assess software engineers’ skills.

Keywords Crowd-sourced design and engineering, Task-oriented crowdsourcing, Agile software engineering, Tools and platforms to support crowd science and engineering

Paper type Research paper

1. Introduction
Most hiring managers in software companies understand that a successful member of the software engineering team needs to be strong in both programming skills (e.g. software design, coding and testing skills) and soft skills (e.g. communication, collaboration and self-management skills). Programming skills can be gauged, at least in part, from students’ performance in examinations and programming contests. Soft skills are much harder to assess, especially during the limited time given in job interviews. Although the concept of these skills can be taught, the ability to apply them consistently in practice can only be acquired through one’s own experience.

In tertiary education institutions, software engineering students are often assessed by a combination of examinations and coursework projects. Many educators have realized the limitations of examinations in assessing students’ practical skills. Thus, coursework projects often serve as an opportunity for students to both practice and demonstrate their skills. However, as an instructor has to face tens or even hundreds of students in a semester, it is not practical for him or her to know the weaknesses and strengths in each student’s skills in detail through observation. Technologies that can subjectively quantify students’ skill development are needed to enable instructors to proactively and effectively help each student.

With the emergence of systems capable of collecting personal behaviour trajectory big data (Heymann and Garcia-Molina, 2011), data-driven analysis of people’s characteristics over time is changing how students’ performance can be measured. Some funding agencies are starting to support research in data-driven student assessment technologies to complement traditional examination scores. For example, the Ministry of Education in Singapore has started an initiative to build technological solutions capable of holistically assessing students’ twenty-first-century competencies (e.g. critical thinking and self-directed learning skills).

Following a similar line of thinking, in this paper, we explore how software development behaviour data can be used to assess students’ programming and soft skills. As agile software development (ASD) involves many human factors reflecting developers’ personal characteristics compared with other plan-driven methodologies (Cockburn and Highsmith, 2001), we focus on tracking students’ activities in the ASD process. For this purpose, we conducted a 12-week study involving 125 undergraduate software engineering students from the Beihang University, Beijing, China. The students self-organized into 21 ASD teams of five to seven persons. Each team developed one software system of significant complexity following the Scrum ASD method as part of their course requirements. Some examples of the coursework projects include “A Personal Healthy Living App”, “A Social Network App for Senior Citizens” and “An Activity Tracking App for the Elderly”.

Students in this study carry out software engineering activities at various stages of the Scrum methodology in our online agile project management (APM) tool – the Human-centred Agile Software Engineering (HASE) platform (www.linjun.net.cn/hase/) (Lin et al., 2014). HASE mainly supports activities during the sprint planning and sprint review/retrospective phases. Such activities include proposing tasks; estimating the priority,
difficulty and time required for each task; deciding how to allocate tasks; gathering collaboration information; reviewing the timeliness and quality of completed tasks; and providing feedback on individual team member’s mood at different points in time during a sprint. During the study, students logged 169,137 ASD activities in the HASE platform. By analysing the collected data set to reflect students’ programming skills, collaboration and mood stability, we demonstrate the potential of this research direction and discuss its implications for software engineering education.

2. Related work
To the best of our knowledge, there has yet to be published previous studies using software engineering activity data to assess software engineers’ skills. Nevertheless, as the skills assessment has always been an important problem, other methods have been applied in an attempt to address it.

In 2004, Kitchenham et al. advocated evidence-based software engineering (EBSE) similar to what is happening with evidence-based medicine (Kitchenham et al., 2004). A technological platform for tracking and analysing important factors in software engineering such as skills factors and life-cycle factors were called for, and the benefits of which were analysed. Nevertheless, the work intends to produce methods to support the development of high-quality software through objective analysis of performance-related indicators. Although similar in principle to our work, they do not specify how EBSE can assess the skills demonstrated by the software engineers or how such insight can be used to improve software engineering education.

In 2011, Salleh et al. presented the results of a systematic literature review concerning agile pair programming effectiveness (Salleh et al., 2011). The paper analysed compatibility factors, such as the feel good, personality and skill level factors, and their effect on pair programming effectiveness. Four metrics were used in the analysis:

1. academic performance;
2. technical productivity;
3. program/design quality; and
4. learning satisfaction.

As the study was not focused on assessment, the general findings are not useful for skills assessment. Nevertheless, it did point towards the importance of soft skills in software engineering.

In 2014, Lin et al. started to track personal performance data with APM tools to study task allocation-related decision-making under Scrum (Lin et al., 2014). It used the same research techniques as reported in this paper. However, the study focused on analysing students’ programming skills and did not consider their soft skills such as collaboration and mood.

3. Study design
In this section, we present our research approach and the metrics adopted in our analysis.

3.1 Research approach
We use the HASE APM platform to unobtrusively track the student participants’ activities in the Scrum ASD process, including their decision-making, collaboration, task assignment and mood. The platform provides six main features to support APM that cover the sprint planning and sprint review/retrospective phases:
(1) **Registration:** To build user profiles, HASE requires registrants to specify their self-assessed competence levels in different areas of expertise such as familiarity with specific programming languages, system design methodologies and user interface (UI) design tools. This information will only be used to compute an initial assessment for a user in the absence of peer ratings or performance data. Once data from these relatively more objective sources become available, the user’s self-assessment will be excluded from the assessment result.

(2) **Team and role management:** HASE supports the creation of teams, the selection of product owners and stakeholders into the teams and the assignment of different roles within a team (e.g. programmers and UI designers).

(3) **Task management:** Task information including task description, skills required for the task and the person who proposed each task is displayed for all team members to view. The difficulty value of each task $T$ is recorded using an 11-point Likert scale (Likert, 1932) (with 0 denoting “extremely easy” and 10 denoting “extremely hard”). Each team member can input his or her estimated difficulty value for each task into the HASE platform. The HASE platform then uses the average difficulty value for the task ($D_T$). The students were asked to take into account the technical challenge as well as the amount of effort required when judging the difficulty of a task. The priority value of each task is also recorded using an 11-point Likert scale (with 0 denoting “extremely low priority” and 10 denoting “extremely high priority”). Each team member can input his or her estimated priority value for each task into the HASE platform. The HASE platform then uses the average priority value for the task.

(4) **Sprint planning:** HASE records the teams’ decisions on which tasks are assigned to which team member during each sprint. Once assigned, the status of the task becomes “Assigned”. The assignee $i$ inputs his or her confidence value ($Conf_T^i$) for each task $T$ on an 11-point Likert scale (with 0 denoting “not confident at all” and 10 denoting “extremely confident”). Each team member also inputs the estimated required time to complete each task (in number of days). The HASE platform uses the average estimated time required to generate the deadline for the task ($T_{est}^T$). Apart from a primary assignee, multiple students can collaboratively work on a task. The collaborator information for each task is also recorded by HASE.

(5) **Sprint review/retrospective:** Once a task is completed, the assignee changes its status in the HASE platform to “Completed”. This action will trigger HASE to record the actual number of days ($T_{act}$) used to complete this task. HASE also provides functions for team members to peer review the quality ($Qual_T$) of each completed task $T$. The quality of a completed task is recorded in the platform using an 11-point Likert scale, with 0 representing “extremely low quality” and 10 representing “extremely high quality”. The average quality rating for each task is used by HASE as the final quality rating for that task.

(6) **Team morale monitoring:** During the sprint planning meeting, team members can report their current mood values into the HASE platform. A person $i$’s mood at the beginning of a sprint $t$ is represented on a five-point Likert scale, with 1 representing “very low” and 5 representing “very high”. During the sprinter view/retrospective meeting, each task assignee $i$ can report his or her mood after
completing a task at the end of a sprint \( t(m^\text{end}_i(t)) \) using the same five-point Likert scale.

The input data to the HASE platform required from ASD teams are as a result of students’ activities following the Scrum methodology. In this way, users of HASE can behave as if they are using any APM tool without expending additional effort to help with data collection. Thus, the data collection process remains unobtrusive to the participants. Over the 12-week period of the coursework project, the HASE platform collected 169,137 behaviour trajectory records related to software engineering activities from the 125 students who participated in this study.

3.2 Metrics
In this paper, we adopt the exploratory data analysis (EDA) approach (Tukey, 1977) to analyse the data collected. EDA is an approach for analysing data sets to summarize their main characteristics, often with visual methods. It is primarily for understanding what can be learnt from the data beyond the formal modelling or hypothesis testing task. We use the following metrics to facilitate our analysis:

- **Technical productivity** \( \mu_i \): It refers to the average amount of workload a student \( i \) can complete during a sprint. In this study, we use the task difficulty value as an indicator of the workload of a task, as the task difficulty values reported by students denote both the technical challenge and the amount of effort required to complete the task.

- **Competence** \( \text{Comp}_i \): It refers to the probability a student \( i \) can complete a task assigned to him or her with satisfactory quality before the stipulated deadline. In this paper, the outcome of a task needs to achieve an average quality rating higher than five out of ten to be considered as having satisfactory quality. This metric is similar to a student’s reputation. Thus, we adopt a reputation computation model – the beta reputation model (Jøsang et al., 2007) – which is widely used in the fields of online services, artificial intelligence and network communications (Pan et al., 2009; Yu et al., 2010, 2011; Liu et al., 2013; Yu et al., 2013a). It is calculated as follows:

\[
\text{Comp}_i = \frac{\alpha_i + 1}{(\alpha_i + 1) + (\beta_i + 1)} (1)
\]

where \( \alpha_i \) and \( \beta_i \) are calculated as:

\[
\alpha_i = \sum_{T \in \mathcal{G}(i)} 1(T^\text{eff}_T - T^\text{est}_T \leq 0 \text{ and } \text{Qual}_T > 5) \mathcal{D}_T (2)
\]

\[
\beta_i = \sum_{T \in \mathcal{G}(i)} 1(T^\text{eff}_T - T^\text{est}_T > 0 \text{ and } \text{Qual}_T \leq 5) \mathcal{D}_T (3)
\]

The function \( 1[\text{condition}] \) in equation (2) and equation (3) equals to 1 if “condition” is true. Otherwise, \( 1[\text{condition}] \) equals to 0. \( \mathcal{G}(i) \) denotes the set of tasks \( i \) has previously worked on until the current point in time. The “+1” terms in the numerator and denominator of equation (1) are Laplace smoothing terms (Wang and Singh, 2007) which ensure that if \( i \)
has no previous track record, Comp\textsubscript{i} evaluates to 0.5, indicating maximum uncertainty about i’s performance:

- **Team morale (begin)** (M\textsubscript{\text{begin}}\textsubscript{i}(t)): It refers to the average of the mood values reported by members of team j during the sprint planning meeting of sprint t.
- **Team morale (end)** (M\textsubscript{\text{end}}\textsubscript{i}(t)): It refers to the average of the mood values reported by members of team j during the sprint review/retrospective meeting of sprint t.

4. Results and analysis

An EDA has identified certain personal characteristics that may become useful markers for assessing students’ skills in the future. Figure 1 shows the participants’ competence scores versus their productivity scores at the end of the study. It can be observed that the participants’ performances in terms of these metrics are quite distinguishable. In general, participants who demonstrated high competence tend to also be able to handle high workloads allocated to them ($r = 0.7443$, $p < 0.01$). One participant achieved significantly higher competence and productivity scores than the rest of the participants.

Collaboration is generally regarded as a useful way to improve the effectiveness and efficiency of a software team. Figure 2 shows a heat map of the number of collaborators per task each participant had for each of the 12 weeks. The lighter the colour of a point on the figure, the more collaborators per task that the participant had for that particular week. The colour scale mapping different colour gradients to the actual number of collaborators per task is shown on the right-hand side of the figure. Participants are ranked according to their average number of collaborators per task per week. Those who are shown at the bottom of the figure ranked the highest among their peers. It can be observed that this metric can distinguish the behaviours among different participants clearly.

Stability of mood is a sign showing one’s maturity and self-management skills. Figure 3 shows a heat map of the intra-week mood change (which is computed as $\Delta m_i(t) = m_i^{\text{end}}(t) - m_i^{\text{begin}}(t) \in (-5, 5)$ for each week) over the 12 weeks. In all, 102 out of the 125 participants provided valid reports on their $m_i^{\text{begin}}(t)$ and $m_i^{\text{end}}(t)$ values. The colour
scale mapping different colour gradients to the intra-week mood change is shown on the right-hand side of the figure. Participants are ranked according to their average intra-week mood change values per week over 12 weeks. Those who are shown at the bottom of the figure ranked the highest among their peers. It can be observed that this metric can distinguish the behaviours among different participants quite clearly. The mood of those who ranked high on this metric tends to increase at the end of a week after a sprint of development. And as their mood at the beginning of the week also tends to be high, the increments are generally small. Thus, their mood remain relatively stable throughout a sprint. Those who ranked low on this metric (top part of the figure) tend to have big negative mood swings, especially towards the end of the study.
To explore if the assessment of participants’ skills may help us identify students who are good at hands-on software engineering but did not stand out in examinations, we construct a skills score to aggregate the effect of competence, productivity, collaboration and mood stability into one scalar measurement. In this study, the skills score, $S_{\text{skills}}(i)$, for a participant $i$ is computed as:

$$S_{\text{skills}} = \frac{S_{\mu_i} + S_{\text{Comp}_i} + S_{\text{col}_i}}{1 - S_{\Delta m_i}} \times \frac{100}{3}$$

where $S_{\mu_i} \in [0, 1]$, $S_{\text{Comp}_i} \in [0, 1]$, $S_{\text{col}_i} \in [0, 1]$ and $S_{\Delta m_i} \in (-1, 1)$ are the normalized scores for $i$ in terms of productivity, competence, collaboration and mood stability, respectively ($S_{\text{skills}}(i) \in [0, 100]$).

Figure 4 plots the participants’ skills scores against their examination scores for the subject of software engineering in the same semester. The examination paper used was the standard software engineering end-of-semester examination paper from the Beihang University, which has been designed by the professors in charge of the course and reviewed by the university examination board. It can be observed that, according to their exam scores, their performance clustered in the range of 80 to 100 marks, enabling almost all of them to achieve a grade of A or A+. However, their skills scores spread from as low as 10 marks to as high as 80 marks, making their performance more distinguishable compared with their exam scores. The skills scores have only a weak positive correlation with the exam scores ($r = 0.2129$, $p < 0.05$). Furthermore, the top three best-performing participants in terms of skills scores achieved only average exam scores among their peers, and many participants with high exam scores achieved low skills scores.

We acknowledge that there may be other ways to compute the skills score and we refrain from claiming that our current formulation for the skills score is the most effective. Nevertheless, the results show that the data-driven skills score can potentially help us distinguish the performance of software engineering students better than examination-based assessments.
Figure 5 shows the distribution of students’ average self-reported mood values during the sprint planning meeting at the start of each sprint. The colour scale represents the average self-reported mood values. The average mood value is 3.86 out of 5. The correlation between students’ mood during the sprint planning meetings and their competence values is $r = 0.0025$, $p < 0.9394$, indicating no statistically significant correlation. The correlation between students’ mood during the sprint planning meetings and their technical productivity values is $r = 0.1505$, $p < 0.01$, indicating a statistically significant albeit weak positive correlation.

Figure 6 shows the distribution of students’ average self-reported mood values during the sprint review/retrospective meeting at the end of each sprint. The colour scale represents the average self-reported mood values. The average mood value is 3.80 out of 5, which is slightly lower than at the beginning of the sprint. The correlation between students’ mood during the sprint review/retrospective meetings and their competence values is $r = 0.0148$, $p < 0.5946$, indicating no statistically significant correlation. The correlation between students’ mood during the sprint review/retrospective meetings and their technical productivity values is $r = 0.4207$, $p < 0.01$, indicating a statistically significant positive correlation. Therefore, based on these analyses, team members with high technical productivity tend to have a high morale, especially at the end of a sprint after completing the tasks allocated to them.

5. Implications

By providing a technological platform for the longitudinal tracking of software engineers’ behaviour trajectory data related to software development, we open up new possibilities for different parties involved in software engineering, namely, researchers, educators and practitioners.

5.1 Implications for software engineering researchers

The availability of large software engineering behaviour data sets will present new challenges to researchers to develop new analytics techniques. With detailed information on
each user’s demographics, skills indicator values over time, detailed interactions with the software engineering tools provided, interactions with other team members and decisions made, the high dimensionality of the data sets makes it a challenge to identify which feature, or combination of features, can form accurate predictors for certain behaviours of interest. Machine learning (Anzai, 1992) can be leveraged to develop useful techniques for this purpose.

However, before this step can happen, additional efforts are needed to complement the behaviour data sets with labelled meta-data on what the observed behaviour patterns mean. This can potentially be achieved by conducting follow-up interview-based studies with the participants through carefully designed questionnaires once unique behaviour patterns have been identified. This also further opens up the research question on how to present the behaviour pattern data in a human-interpretable format to facilitate the interviews.

5.2 Implications for software engineering educators
Software engineer educators may be a viable source of knowledge in the effort of building up a repository of meta-data for the behaviour patterns obtained by the tracking platform. As they frequently interact with students who may be using the proposed tracking platform, they can potentially provide insights into the meanings of the behaviour patterns. The challenge here is for software engineering researchers to provide tools to enable educators who are willing to contribute meta-data for the behaviour patterns to do so with ease. Techniques from the field of crowdsourcing (Doan et al., 2011) may offer a starting point for such an effort.

Once new techniques for automatically assessing a student’s skill development based on his or her behaviour patterns are developed, new forms of real-time personalized inventions may become available to educators. The simplest possibility is for the system to send out alerts on students who may require help in specific areas to course instructors. Through mining the behaviour patterns of many students and cross-checking with their academic performance, or even employment prospects if such data are available, the system may be able to suggest behaviour trajectories that are the most beneficial for students from different backgrounds, thereby making data-driven personalized software engineering training
possible. The envisioned behaviour data tracking platform can potentially convert software engineering education into a test bench for open science and enable a more adaptive and individualized learning experience.

5.3 Implications for software engineering practitioners
The behaviour data are tracked in an unobtrusive manner by the APM tool automatically. The peer review-related functions represent activities that an ASD team member already has to perform when following the ASD practice. Overall, the proposed APM tool-based behaviour data tracking approach does not require software engineers to incur additional overhead. However, the data analytics functions provide ASD teams with insights into detailed team dynamics and performance information that can be useful for decision-making. Furthermore, with the behaviour data as input, automatic context-aware software engineering task allocation decision support mechanisms (Lin, 2013) become a distinct possibility. These mechanisms can be based on similar mechanisms available in the field of crowdsourcing (Yu et al., 2013b, 2013c, 2015).

6. Discussions and future work
In this paper, we explore a novel data-driven approach to assess software engineering students’ skills. Different from traditional interview/internship-based methods, our study is based on participants’ ASD activity trajectory data collected unobtrusively during normal ASD processes through our HASE APM platform. This type of data objectively reflects developers’ ASD activities and performance at fine granularity.

As the data collection and analytics technologies further develop, software engineering students may eventually perform all coursework activities in a technology platform capable of unobtrusively collecting their behaviour data and continuously assessing a wide range of their skills over time. In this way, the students’ practical skill development can be monitored by their instructors so that pedagogical methods can be personalized to help individual students in specific areas. Such a tool will enable software engineering educators to have a quantifiable way of understanding their students’ skill development and take a proactive approach in helping them develop programming and soft skills. The skills scores may, one day, be part of a student’s academic profile and be taken into consideration by industry recruiters to help companies identify well-rounded software engineering talents suitable for their teams.

From this study, we see the start of a series of research and applications in data-driven software engineering skills assessment. In future research, we plan to conduct surveys/interviews to understand more in-depth how students collaborate. We will continue using the HASE platform to collect agile programming activity data over subsequent semesters and expand our data collection effort to include more universities so as to investigate the possible effects of socio-cultural factors. More finely grained data such as the time each student spent on a task and the breakdown of the usage of the time will also be collected in future versions of the HASE platform.

References


Further reading
About the authors

Jun Lin is a Research Fellow at the Joint NTU-UBC LILY Research Centre, Nanyang Technological University, Singapore; a Research Scientist at the Alibaba-NTU Singapore Joint Research Institute; and an Adjunct Associate Professor at College of Software, Beihang University, China. He holds a PhD degree from the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His current research interests include LoRaWAN, blockchain, Internet of Things, crowd science, software engineering and AI technologies. Jun Lin is the corresponding author and can be contacted at: junlin@ntu.edu.sg

Han Yu is a Nanyang Assistant Professor (NAP) at the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He has been a Visiting Scholar at the Department of Computer Science and Engineering, Hong Kong University of Science and Technology (HKUST) from 2017 to 2018. Between 2015 and 2018, he held the prestigious Lee Kuan Yew Post-Doctoral Fellowship (LKY PDF) at the Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY). His research focuses on artificial intelligence-powered crowd-computing.

Zhengxiang Pan is a PhD student at Interdisciplinary Graduate School, Nanyang Technological University, Singapore. His research focuses on human–computer interaction, crowdsourcing and social mobilization.

Zhiqi Shen is a Senior Scientist at the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He obtained his BSc degree in Computer Science and Technology from Peking University, MEng in Computer Engineering from Beijing University of Technology and PhD from Nanyang Technological University. His research interests include artificial intelligence, software agents, multi-agent systems, goal-oriented modelling, agent-oriented software engineering, semantic web/grid, e-Learning, bioinformatics and bio-manufacturing, agent augmented interactive media, game design and interactive storytelling.

Lizhen Cui is a Professor at the School of Computer Science and Technology, Shandong University, China. His research interests include big data management and big data analytics, big data artificial intelligence, service computing and collaborative computing, software architecture and technology in cloud.

For instructions on how to order reprints of this article, please visit our website: www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: permissions@emeraldinsight.com
A scientometric analysis of e-participation research

Tuotuo Qi, Tianmei Wang, Yanlin Ma and Wei Zhang
School of Information, Central University of Finance and Economics, Beijing, China, and
Yanchun Zhu
Business School, Beijing Normal University, Beijing, China

Abstract

Purpose – Due to the increasing demand for public services, as a new form of public governance, e-participation has emerged. Scholars from various disciplines have published plenty of research results on e-participation. This paper aims to reveal the research status frontiers directly by mapping knowledge domains.

Design/methodology/approach – The authors take 1,322 articles on e-participation published in Web of Science from 2001 to 2017 as research object. They then run the information visualization software CiteSpace to drill deeper into the literature data.

Findings – The study found that e-participation research has the obvious interdisciplinary feature; the author and institution cooperation networks with less internal cooperation are relatively sparse; the USA ranks first in the field of e-participation research, followed by the UK, with the other countries lagged behind; and e-participation through social media is gradually becoming the new research focus.

Originality/value – Based on the objective data and information visualization technology, the research intuitively reveals the research status and development trend of e-participation.

Keywords Social media, E-participation, Citation bursts, Co-occurrence network, Cooperation network, Time-zone

Paper type Literature review

1. Introduction

Since the 1990s, the internet has continuously brought about tremendous changes in human life. In the real world, internet has been integrated into government affairs. For example, the USA took the lead in e-government. In 1993, the Clinton Administration established the “National Performance Evaluation Committee” and proposed to improve public service through e-government. In 2016, 120 countries out of 193 member states have developed electronic decision tools (Nations, 2016).

The increasing popularity of e-participation has attracted more and more scholars to devote themselves into the e-participation research. Citizens participated in environmental...
governance through GeoTools or EoC platform (Kingston et al., 2000; Kingston, 2008). After the earthquake and tsunami, Onagawa in Japan implemented an online participatory mechanism (Aoki, 2017). The online and offline participation models of budget system in Guangzhou, South Korea, had different impacts on decision-making (Lim and Oh, 2016). MD Robbins conducted a web-based survey at the town of West Hartford which allowed real-time interaction, asking respondents to weigh options between service level and tax amounts to help decision makers better understand citizen preferences (Robbins et al., 2008).

Patrick Dunleavy proposed the concept of digital governance, emphasizing the importance of e-government and network participation (Dunleavy et al., 2006). The development of e-participation is divided into four stages, that is, the “bulletin board” stage, “partial service delivery” stage, “portal with secure operability and integral service” and “interactive democracy” stage (West, 2011). This clearly indicates that democracy and civil rights are the future trend of e-government. Some scholars use comparative research method to analyze the status of e-participation in various countries. Andrew Chadwick studied the online political operational mechanism in countries such as the USA and UK and demonstrated the impact of internet on local democracy, social movements, and elections (Chadwick, 2010). Nahleen Ahmed analyzed the status quo of e-government in the USA, UK, Singapore, Canada and other countries based on information acquisition, online services and interaction and then pointed out their respective advantages and disadvantages (Ahmed, 2006).

The e-participation creates a dialogue between the public and the government. Through studying e-participation systematically and comprehensively, we can know the ways in which government improve the form, channel and process of e-participation. Finally, it distinctly helps the government understand what the public is thinking and how to put it into practice. So, it is valuable to reveal the status and frontiers of e-participation.

To more fully and objectively grasp the status quo and refine research hotspots of e-participation research, we use CiteSpace to analyze the e-participation research literature retrieved from the Web of Science Core Collection database from 2001 to 2017.

2. Method and data collections

2.1 Method
Visual analysis of citations is an important branch of information visualization. It first processes a large amount of citation data and then uses information visualization technology to make it easier for people to find hidden patterns (Yunjing and Hanqing, 2007). To more fully grasp the development of e-participation research, we selected CiteSpace for visual analysis. CiteSpace is a multi-dimensional, time-sharing and dynamic information visualization tool developed by Professor Chen Chaomei of Drexel University in the USA. It has the analysis functions of cooperation network co-occurrence network and burst detection function which can be used to identify new research hotspots and frontiers (Yue et al., 2015; Chen, 2006). With the continuous optimization of algorithms and functions, CiteSpace has been widely used in more than 60 fields such as computer science, information science, and medical science (Jie and Chaomei, 2016). We use CiteSpace to conduct data mining and econometric analysis of e-participation research for grasping the evolutionary path and future trends.

2.2 Data collection
Integration of information and communication technologies (ICTs) into various fields of government affairs has been a topic of discussion. E-government refers to the use of ICTs to improve public service delivery, and that e-democracy is the follow-up phase of e-government.
which refers to the use of ICTs to increase the participation degree of democratic governance (Gunter, 2006). E-government is initiated by government, whereas e-democracy involves multiple stakeholder initiatives (Gowda and Gupta, 2010). Therefore, e-government is the first step toward online democratic participation (Ingram and Smith, 1993). However, some scholars have divided the development of the e-government into four phases, namely, the “bulletin board” stage, “partial service delivery” stage, “portal with secure operability and integral services” and “interaction democratic” stage (West, 2011). We can see that only the fourth stage of “interactive democracy” includes citizens’ participation. Therefore, e-government has a broader meaning than e-democracy. Then e-participation refers to support democratic decision-making by promoting deliberation between government and citizens through ICTs (Macintosh, 2004). J Millard criticized the definition and argued a more “administrative” perspective should be adopted, noting that e-participation is not only limited to the use of ICTs for democratic decision-making but also involves a series of processes that do not directly involve democratic decision-making but still use ICTs for participation (Millard, 2009). In this sense, e-participation belongs to the fourth stage of e-government.

However, many previous literature reviews on e-participation did not take into account the importance of participation resulting that e-government is equated with e-participation when be used in searching literatures (Zolotov et al., 2018; Rodriguez-Bolivar et al., 2018). Considering the inclusion relation between e-government and e-participation, we integrate e-government with participation; that is to equate “e-government AND participation” with “e-participation.” Besides, many literature reviews did not take “citizen engagement” into consideration. However, we find that the literature with retrievable field containing “citizen engagement”, not including keywords such as “e-government”, “e-democracy,” is still related to the online deliberation between government and citizens (Medaglia and Zhu, 2017; Tettey, 2017; Ertiö and Bhagwatwar, 2017). Therefore, to ensure the authority of retrieved literatures and high recall ratio of the search query as much as possible, the following retrieval type (Table I) is adopted.

The data were retrieved from the Web of Science Core Collection database. Each data record mainly includes authors, title, abstract, keywords and citations. In consideration of the fact that the literatures published in 2018 have not yet been fully included in WoS, we

<table>
<thead>
<tr>
<th>Keywords 1</th>
<th>Keywords 2</th>
<th>Keywords 3</th>
<th>Keywords 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-participation</td>
<td>E-governance AND participation</td>
<td>Citizen engagement AND internet</td>
<td>E-democracy</td>
</tr>
<tr>
<td>Electronic-participation</td>
<td>Electronic-governance AND participation</td>
<td>Citizen engagement AND social media</td>
<td>Electronic-democracy</td>
</tr>
<tr>
<td></td>
<td>E-government AND participation</td>
<td>Citizen engagement AND information system</td>
<td>Democracy AND internet</td>
</tr>
<tr>
<td></td>
<td>Electronic-government AND participation</td>
<td>Citizen engagement AND digital</td>
<td>Democracy AND social media</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Democracy AND information system</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Democracy AND digital</td>
</tr>
</tbody>
</table>

**Notes:** TS = (“e-participation” OR “electronic-participation” OR (“e-governance” AND participation) OR (“electronic-governance” AND participation) OR (“e-government” AND participation) OR (“electronic-government” AND participation) OR “e-democracy” OR “electronic-democracy” OR (democracy AND (internet OR social media OR “information system” OR digital)) OR (“citizen engagement” AND (“social media” OR digital OR internet OR information system”)))
chose December 31, 2017, as the deadline. Time span was set “from 2001 to 2017”. The language type was “English” and the document type was “ARTICLE”. Based on the advanced search function of WoS, the retrieval type was continuously optimized and adjusted according to the search results. We obtained 1,367 literatures studies through the initial retrieval. After the non-related literatures including forms such as “i.e. participation” were removed 1,322 document data were finally obtained.

3. Basic distribution analysis
3.1 The number of annual publications distribution analysis
The number of annual publications is an important index for measuring the development of scientific research, as it reflects, to a certain extent, the changes in knowledge quantity. Therefore, the distribution of annual publications (Figure 1) can form a preliminary understanding of e-participation research. It can be seen that in 2014, the number of publications increases rapidly indicating e-participation began to receive high attention from academia. However, the number of publications remains nearly constant since 2015. This manifests that the scholars’ research focus has stabilized, so it is necessary to summarize the previous research hotspots and then find breakthroughs to prepare for further research. On the whole, the number of publications has grown steadily from 15 in 2001 to 156 in 2017.

3.2 Journal distribution analysis
Journal analysis provides guidance for scholars to select platforms for data collection and publishing their research. The 1,322 selected studies are distributed in 447 core academic journals. Overall, the distribution of research literature is scattered. However, as can be seen from Table II, there are 20 academic journals with a total of ten or more publications which add up to 472, accounting for 35.7 per cent of the total. Among them, “Government Information Quarterly” tops the list with 88 publications, manifesting the journal greatly favors the subject of e-participation. “Information Communication Society” and “New Media Society” have issued more than 50 articles. In addition, journals with an impact factor greater than 2 account for 50 per cent of the 20 journals which proves there is a high possibility of publishing research papers on e-participation in high-level journals. It also proves that high-level research scholars and journals have begun to pay attention to e-participation, thus reflecting a certain degree of maturity in this field. In addition, the journals marked with *** and ** in the table are core magazines to which researchers should refer.
3.3 Subject distribution analysis

Table III shows that e-participation research is interdisciplinary. The “Communication” discipline has the largest publication number, accounting for 30.33 per cent of the total which demonstrates e-participation mainly involves the interaction between government and citizens. “Information Science Library Science” ranks second with 259 publications, and “Computer Science” ranks fourth with 174 publications, indicating the importance of information technology in the field of e-participation. “Government Law” ranks third and “Public Administration” ranks sixth, both with the publication number more than 100. “Sociology” and “Social Sciences Other Topics” rank fifth and seventh, respectively, mainly because of the universality of sociological studies, including micro-level social actions and human interactions, as well as macro-level social systems and structures. The number of

<table>
<thead>
<tr>
<th>Rank</th>
<th>Subject categories</th>
<th>Publication no.</th>
<th>(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Communication</td>
<td>401</td>
<td>30.33</td>
</tr>
<tr>
<td>2</td>
<td>Information Science Library Science</td>
<td>259</td>
<td>19.59</td>
</tr>
<tr>
<td>3</td>
<td>Government Law</td>
<td>200</td>
<td>15.13</td>
</tr>
<tr>
<td>4</td>
<td>Computer Science</td>
<td>174</td>
<td>13.16</td>
</tr>
<tr>
<td>5</td>
<td>Sociology</td>
<td>130</td>
<td>9.83</td>
</tr>
<tr>
<td>6</td>
<td>Public Administration</td>
<td>115</td>
<td>8.70</td>
</tr>
<tr>
<td>7</td>
<td>Social Sciences Other Topics</td>
<td>92</td>
<td>6.96</td>
</tr>
<tr>
<td>8</td>
<td>Business Economics</td>
<td>89</td>
<td>6.73</td>
</tr>
<tr>
<td>9</td>
<td>Area Studies</td>
<td>43</td>
<td>3.25</td>
</tr>
<tr>
<td>10</td>
<td>Psychology</td>
<td>37</td>
<td>2.80</td>
</tr>
</tbody>
</table>

Table III.
The top ten highly cited subject categories in e-participation area

Notes: The *** in the table indicates that journals with an impact factor greater than 3 are the most important; ** indicates that journals with an impact factor greater than 2 are very important; * indicates that journals with an impact factor greater than 1 are important.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Journal title</th>
<th>Count</th>
<th>(%)</th>
<th>IF 2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Government Information Quarterly**</td>
<td>88</td>
<td>6.70</td>
<td>4.09</td>
</tr>
<tr>
<td>2</td>
<td>Information Communication Society***</td>
<td>53</td>
<td>4.00</td>
<td>2.692</td>
</tr>
<tr>
<td>3</td>
<td>New Media Society***</td>
<td>51</td>
<td>3.90</td>
<td>4.18</td>
</tr>
<tr>
<td>4</td>
<td>International Journal of Communication*</td>
<td>30</td>
<td>2.30</td>
<td>1.498</td>
</tr>
<tr>
<td>5</td>
<td>Javnost-the Public</td>
<td>27</td>
<td>2.00</td>
<td>0.413</td>
</tr>
<tr>
<td>6</td>
<td>Social Science Computer Review</td>
<td>24</td>
<td>1.80</td>
<td>2.293</td>
</tr>
<tr>
<td>7</td>
<td>Electronic Government Proceedings</td>
<td>22</td>
<td>1.70</td>
<td>0.402</td>
</tr>
<tr>
<td>8</td>
<td>Media Culture Society*</td>
<td>20</td>
<td>1.50</td>
<td>1.128</td>
</tr>
<tr>
<td>9</td>
<td>Telematics and Informatics***</td>
<td>19</td>
<td>1.40</td>
<td>3.398</td>
</tr>
<tr>
<td>10</td>
<td>International Journal of Press-Politics*</td>
<td>17</td>
<td>1.30</td>
<td>1.523</td>
</tr>
<tr>
<td>11</td>
<td>Information Society*</td>
<td>15</td>
<td>1.13</td>
<td>1.558</td>
</tr>
<tr>
<td>12</td>
<td>Computers in Human Behavior***</td>
<td>13</td>
<td>0.98</td>
<td>3.435</td>
</tr>
<tr>
<td>13</td>
<td>Journalism*</td>
<td>13</td>
<td>0.98</td>
<td>1.484</td>
</tr>
<tr>
<td>14</td>
<td>Political Communication**</td>
<td>13</td>
<td>0.92</td>
<td>2.467</td>
</tr>
<tr>
<td>15</td>
<td>Public Administration Review***</td>
<td>13</td>
<td>0.92</td>
<td>3.473</td>
</tr>
<tr>
<td>16</td>
<td>Electronic Government Proceedings</td>
<td>12</td>
<td>0.91</td>
<td>4.02</td>
</tr>
<tr>
<td>17</td>
<td>American Behavioral Scientist*</td>
<td>11</td>
<td>0.83</td>
<td>1.311</td>
</tr>
<tr>
<td>18</td>
<td>Journal of Communication***</td>
<td>11</td>
<td>0.83</td>
<td>3.914</td>
</tr>
<tr>
<td>19</td>
<td>Communication Research***</td>
<td>10</td>
<td>0.76</td>
<td>3.021</td>
</tr>
<tr>
<td>20</td>
<td>European Journal of Communication*</td>
<td>10</td>
<td>0.76</td>
<td>1.408</td>
</tr>
</tbody>
</table>

Notes: The *** in the table indicates that journals with an impact factor greater than 3 are the most important; ** indicates that journals with an impact factor greater than 2 are very important; * indicates that journals with an impact factor greater than 1 are important.
“Area Studies” reaches 43, indicating that different regions may have different forms and development degrees of e-participation, thus attracting the attention of scholars in this field. In particular, it is worth noting that “Business Economics” and “Psychology” have entered the top ten, demonstrating that e-participation also involves the macro-development of business economy and micro-psychological factors. Therefore, e-participation research is an interdisciplinary field and it is extremely significant to strengthen the communication and cooperation among scholars in various fields.

4. Keywords analysis

4.1 Analyze the time-zone of keywords

The time that the node was first cited determines its time-zone position, making time-zone map take the evolution of knowledge as its core content (Yue, 2014; Jie and Chaomei, 2016). By plotting the time zone of keywords, we can obtain the dynamic process of the e-participation research from overall to each part (see Figure 2). In 2001, “democracy” has emerged as a keyword, indicating that the play of democratic nature through internet dates back to 2001 at the latest. In 2002, “communication” appeared, implying e-participation’s main purpose is to facilitate communication between government and citizens. In 2003, “United States” appeared, which means that the USA takes the top spot in the field of e-participation. In contrast, “China” as a keyword began to appear in 2011. “E-government” first appeared in 2005, whereas “e-participation” appeared in 2008, further proving that e-participation belongs to the “interactive democracy” stage of e-government. In 2010, “social media” appeared for the first time as a keyword, indicating that the public has begun to participate in government administration through social media. Immediately thereafter, “Facebook” appeared in 2012 and “Twitter” began to appear in 2013.

4.2 Co-occurrence network and citation bursts of keywords

Keywords, as the core and entry point of an article, concisely summarize the research content. Combining Table IV and Figure 3, we find “democracy” cited 388 times is the largest node in the network, followed by “internet” cited 369 times and “social media” cited 196 times. But none of their centrality values enter the top ten. “Communication” with the centrality value of 0.07 is in the middle of the network which indicates the main purpose of e-participation is to strengthen the communication and continuously increase the degree of public participation in social governance. Then “United States,” “community,” “trust” and

Figure 2.
The time-zone of keywords in e-participation area
“civil society” have a centrality value of 0.06, reflecting their important role in e-participation research.

The central node analysis can explore the key works, whereas burst detection is more suitable for detecting emerging trends and sudden changes in subject development (Chaomei et al., 2009; Jie and Chaomei, 2016). Citation burst refers to the sudden increase of keyword in the literature. The citation bursts are sorted by starting time (Table IV). From 2009 to 2011, “web” and “media” began to emerge, indicating the government increasingly uses media and network to provide social governance channels for citizens. In 2012, “service” began to appear, which means that the level of e-participation has developed from simple information informing to the provision of complaint channels and consulting services. Between 2015 and 2017, keywords such as “journalism”, “deliberation” and “twitter” emerged, showing that in the era of big data, forwarding news and deliberation network on Twitter have become research priorities. In addition, “web” and “media” lasted only a short time, indicating they did not receive continued attention from researchers. The likely reason is that scholars have begun to focus on more detailed research scenarios such as “Twitter”, rather than “Web” and “media” whose meanings are too broad. It can be found that most of hot topics in social media are related to hard news (Lawrence et al., 2013). About 63 per cent of users get news from social media (Newman et al., 2015). Journalists speed up the flow of information and break the elite monopoly on agenda-setting, mobilizing citizen participation and political participation through social media platforms (Hermida et al., 2014). In addition, media agencies mainly use twitter as information source and brand promotion channels (Lasorsa et al., 2012).

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Citation counts</th>
<th>Keywords</th>
<th>Centrality</th>
<th>Bursts</th>
<th>Begin</th>
<th>End</th>
<th>Strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democracy</td>
<td>388</td>
<td>Communication</td>
<td>0.07</td>
<td>Twitter</td>
<td>2015</td>
<td>2017</td>
<td>2.967</td>
</tr>
<tr>
<td>Internet</td>
<td>369</td>
<td>United states</td>
<td>0.06</td>
<td>Deliberation</td>
<td>2015</td>
<td>2017</td>
<td>4.0135</td>
</tr>
<tr>
<td>Social media</td>
<td>196</td>
<td>Community</td>
<td>0.06</td>
<td>Journalism</td>
<td>2015</td>
<td>2017</td>
<td>4.5522</td>
</tr>
<tr>
<td>Participation</td>
<td>181</td>
<td>Trust</td>
<td>0.06</td>
<td>Service</td>
<td>2012</td>
<td>2013</td>
<td>4.5515</td>
</tr>
<tr>
<td>Media</td>
<td>120</td>
<td>Civil society</td>
<td>0.06</td>
<td>Website</td>
<td>2011</td>
<td>2012</td>
<td>3.383</td>
</tr>
<tr>
<td>Communication</td>
<td>119</td>
<td>E-government</td>
<td>0.05</td>
<td>Opinion</td>
<td>2010</td>
<td>2013</td>
<td>3.285</td>
</tr>
<tr>
<td>E-government</td>
<td>111</td>
<td>News</td>
<td>0.05</td>
<td>Web 2.0</td>
<td>2010</td>
<td>2013</td>
<td>3.816</td>
</tr>
<tr>
<td>Information</td>
<td>105</td>
<td>Policy</td>
<td>0.05</td>
<td>Media</td>
<td>2009</td>
<td>2010</td>
<td>3.0586</td>
</tr>
<tr>
<td>E-democracy</td>
<td>90</td>
<td>Deliberative</td>
<td>0.05</td>
<td>Mass media</td>
<td>2009</td>
<td>2012</td>
<td>3.9222</td>
</tr>
<tr>
<td>political participation</td>
<td>85</td>
<td>Management</td>
<td>0.05</td>
<td>Web site</td>
<td>2009</td>
<td>2011</td>
<td>4.0679</td>
</tr>
</tbody>
</table>

Figure 3.
Co-occurrence network of keywords in e-participation area
Therefore, through keyword analysis, we find that the main purpose of e-participation is to promote communication and strengthen the level of democratic governance; the USA is more advanced in e-participation and thus receives research scholars’ attention and the citizens have begun to participate in government administration through social media, especially Twitter.

5. Authors, institutions and countries analysis

5.1 The author cooperation network analysis
The author’s co-signature is a common phenomenon in academia. As shown in Figure 4, the network with a density of 0.0109 has 14 teams. In e-participation research, only some cooperative networks represented by Ernst N, Buchel F, Esser F, Engelser S have formed, whereas most scholars are still in their own state of war. The members collaborate mainly due to the colleagues or teacher-student relationships. In the long run, as benign cooperation and interaction will help improve the overall academic level, it is necessary to further strengthen the construction of academic teams. In addition, we found that the Bolivar MPR and Royo S nodes are the largest, reflecting that the two authors have the greatest influence on e-participation research.

5.2 The institution cooperation network analysis
The institution cooperation network has a total of six teams with a density of 0.0054, which is smaller than the author cooperation network (Figure 5). There is almost no connection among most institutions except an obvious cooperation subnet (Figure 6). It can be found that the institutions in the subnet are almost all famous universities in the USA. In addition to the close cooperation cluster formed within the USA, the entire cooperation network is relatively sparse.
5.3 The country time zone map and cooperation network analysis

5.3.1 Country time zone map analysis. As is shown in Figure 7, the USA has conducted e-participation research in 2001. Since then, many developed countries have accelerated the pace of e-participation research such as The Netherlands, Sweden and Italy. However, compared with developed countries, developing countries such as China, Brazil, Mexico, Turkey and Poland appeared relatively later. This indicates that developed countries give the developing countries a lead in e-participation.

5.3.2 Country cooperation network analysis. The density of country cooperation network is 0.123. From Figure 8 and Table V, it can be found that the USA is located at the core of the
network, with the highest citation frequency and centrality value, indicating the USA is in the lead. England ranks second in terms of cited times and centrality. Although ranking third and seventh respectively in terms of cited times and centrality, Spain still occupies a critical position. However, some countries do not appear in both top-ten lists simultaneously.
namely, Australia and Germany, South Africa and Ireland which means they have less influence in the e-participation field.

6. Conclusions

Different from previous literature review, we focus on “participation”, excluding many articles on the initial stage of e-government, and thus making the results more specific and explicable. In addition, CiteSpace is used to vividly and comprehensively demonstrate the development trend of e-participation research. We also find that “e-participation” has become the most frequently used concept in recent years compared with “e-government”, “e-democracy” and other similar terms. Therefore, it provides reference for the selection of terminology.

The basic distribution reveals that e-participation has aroused wide attention of scholars. Author and institution cooperation networks with less internal cooperation are sparse. The USA ranks first in the field of e-participation. As other countries started late, it is necessary to speed up the e-participation process. Through keywords analysis, we found that “social media” with 196 cited times appeared for the first time as a keyword in 2010, followed by “Facebook” and “twitter” appearing in 2012 and 2013 successively. Simultaneously “Twitter” is a burst term that has emerged in recent years. Therefore, e-participation through social media is gradually becoming the research focus.

This paper has several limitations. First, there is a lack of uniform standards for visual analysis. We can conduct further studies by knowledge domains map (Chen et al., 2008). However, there is no unified conclusion on how to evaluate the validity of the results (Chen et al., 2010). Second, the literature data are incomplete. Although we have considered the related topics as comprehensively as possible, it is still inevitable to leave out some literatures. In view of the two limitations, it is necessary to develop a unified standard for visual analysis and include related literatures as much as possible by continually adding keywords and optimizing retrieval type.

The emergence of social media has lowered the threshold of citizen participation (Rauchfleisch and Schäfer, 2015). For example, the Twitter has become a place where members can share community awareness (Gruzd et al., 2011). Besides, it has been used by social movement organizations and activists for political mobilization (Außerhofer and Maireder, 2013). Weibo offers Internet users an opportunity to express their views on political consequences (Song et al., 2016). Therefore, in recent years, researchers have mainly focused on the expression of emotions and reason, proximity mechanism and implementation effects in social media. For expression of emotions and reason, it mainly focuses on analyzing the reasons for the formation of homogenous groups and probing into Table V.

<table>
<thead>
<tr>
<th>Country</th>
<th>Citation counts</th>
<th>Country</th>
<th>Centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>439</td>
<td>USA</td>
<td>0.61</td>
</tr>
<tr>
<td>England</td>
<td>188</td>
<td>England</td>
<td>0.40</td>
</tr>
<tr>
<td>Spain</td>
<td>77</td>
<td>Italy</td>
<td>0.18</td>
</tr>
<tr>
<td>Australia</td>
<td>69</td>
<td>Canada</td>
<td>0.14</td>
</tr>
<tr>
<td>Canada</td>
<td>62</td>
<td>Peoples R China</td>
<td>0.12</td>
</tr>
<tr>
<td>Peoples R China</td>
<td>46</td>
<td>The Netherlands</td>
<td>0.09</td>
</tr>
<tr>
<td>Germany</td>
<td>43</td>
<td>Spain</td>
<td>0.08</td>
</tr>
<tr>
<td>The Netherlands</td>
<td>43</td>
<td>South Africa</td>
<td>0.08</td>
</tr>
<tr>
<td>Italy</td>
<td>39</td>
<td>Ireland</td>
<td>0.08</td>
</tr>
<tr>
<td>Sweden</td>
<td>36</td>
<td>Sweden</td>
<td>0.07</td>
</tr>
</tbody>
</table>

The top ten countries with the largest citation counts and centrality in e-participation area
the level of interaction between emotions and reason: to judge the feasibility of considering social media as a deliberation forum. For proximity mechanism, groups from different countries or provinces may form their own local network, naturally with heterogeneous characteristics. However, due to the popularity of social media, whether this heterogeneity has been impacted is still unclear. For the implementation effects, it is necessary to further explore the ways and levels of interaction between elite members and citizens. Finally, follow-up studies should focus more on methods selection to go deeper into the process of e-participation.

References


Jie L. and Chaomei C. (2016), Citespace: Technology Text Mining and Visualization, Capital University of economics business press.


Yue C. (2014), The Principle and Application of Citation Space Analysis, Science publishing Company.


Corresponding author
Tuotuo Qi can be contacted at: 642890481@qq.com

For instructions on how to order reprints of this article, please visit our website: www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: permissions@emeraldinsight.com
Abstract
Purpose – This paper aims to clarify the complex path of negative externalities in the sharing economy and proposes corresponding policy recommendations.
Design/methodology/approach – This paper aims to establish an analytical framework for the negative externalities of the sharing economy and to extract the main factors that produce negative externalities, and then, through qualitative comparative analysis method find out how these factors interact to form a negative externality.
Findings – Negative externalities in the sharing economy come from the joint effect of the sharing degree of the product or service and constraint mechanism, and the current main modes of the shared economy increase the possibility of negative externalities.
Originality/value – The paper proposes a complex path resulting from negative externalities in the shared economy.
Keywords Negative externalities, Qualitative comparative analysis, Sharing economy
Paper type Research paper

1. Introduction
The global success of business models represented by Airbnb and Uber have opened various resources for sharing, influencing the rise of sharing economy. The sharing economy has been generated under the background of overcapacity, referring to a new model in which internet companies use algorithms and data to mobilize redundant resources in the market, solve asymmetric market information and improve transaction efficiency. Globally, tens of thousands of various types of sharing economy enterprises exist, affecting people’s lives and consumption patterns in various fields. The 2017 Sharing Economy Development Report released by the China Electronic Commerce Research Center pointed out that in 2017, China’s sharing economy was more than 5tn, with a growth rate of more than 40 per cent since 2016. The sharing economy has played an important role in improving the allocation efficiency and user value of social resources. However, it has also
brought some negative effects, such as shared bikes occupying sidewalks, drivers violating regulations, price increases and tenants destroying homes in short-term rentals.

Recently, countries around the world have begun to actively respond to the negative effects of the sharing economy. For example, in terms of sharing trips, Singapore and other countries recognize the legitimacy of vehicles such as Uber, but under strict regulations; France and others investigate online car rental services as illegal businesses; in California and Colorado in the USA and Washington, DC the legalization of online car rental services has already been achieved. Many studies have begun to focus on governance issues regarding the sharing economy and have proposed various suggested governance methods such as multiple governance (People’s Forum questionnaire survey center, 2017), collaborative governance (Tang, 2017), mixed supervision (Tang, 2015), response-type regulation (Peng, 2016) and so on. However, traditional governance methods produce difficulties for sharing economy as a new, effective economic form. Main reasons for this include the lack of a clear understanding of the emergence of various problems in the sharing economy and the weak pertinence of proposed governance methods. The relative lag in theoretical research has also caused difficulties for relevant departments in the actual governance of the industry.

The sharing economy situation fits well with the externality theory. Pigou (1962) believes that externality is the inevitable result of market operations. When externalities exist, the allocation of resources cannot reach Pareto optimality. Therefore, identifying sources of negative externality may be effective in determining entry points for governance of the sharing economy. Based on the aforementioned, this article discusses the source of negative externalities in the sharing economy to explore which economic features constitute the source of negative externalities and whether specific patterns exacerbate these negative externalities. We further propose corresponding governance strategy.

2. Literature review
The discussion of externalities has been around for a long time, but the concept is still controversial. Historically, the point of view on externalities can be divided into three major nodes. Currently, the general awareness in the academic world attributes the external economy, which was first proposed by Marshall (Zhang, 2012). The external economy, or the changes in business efficiency caused by changes in industry scale, is considered the predecessor of externality. Originally, however, the concept of externality was first put forward by Pigou (1912). In his book, Welfare Economics, the externalities are explained by analyzing the deviation between the marginal private net output value and the marginal social net output value. Pigou believes that externality is actually the inconsistency between marginal private cost and marginal social cost and marginal private income and marginal social income. However, the concept of externality discussed by Coase in the later period was different from Pigou. He proposed that externality is the direct influence of economic activities of the actor on other economic entities in his paper “Problem of Social Cost.” The negative externalities of the sharing economy discussed in this paper are conceptually closer to Coase’s research.

The existence of externalities has been widely accepted; numerous scholars, in addition to discussing the mechanisms, classifications, etc., have also discussed external phenomena of externalities and their appearances (Scitovsky, 1954; Bator, 1958). Externalities have been studied in the banking sector (Su, 2000; Zhang, 2016), insurance (Zhou, 2014), coal mining (Liu, 2014) and other specific industries. Despite the in-depth exploration of externality, it is still a vague concept. In the narrative definition, externality is generally defined as the influence the implementer has on other people in a particular economic activity. For
example, the New Palgrave Dictionary of Economics defines external economies as an effect of a producer's output or input on the nonpayment of another producer. Buchanan and Stubblebine (1962) gave a mathematical representation of externalities: the so-called externality is the independent variable of the welfare function of an economic entity, producing a function that shows if an economic welfare task is affected by other factors not controlled by oneself, there is an externality. Similar definitions include Ping (2006) and Xu (2006). In this type of definition, the subject is clear, but economic activity is ambiguous. The influence of the implementer on other people is an overly broad term. Therefore, key content in the definition is unclear therefore confusing the sources, process of influence and extensions of externalities.

Although the aforementioned representative studies by Marshall, Pigou, Coase and others have greatly improved people's understanding of externality issues, they actually provide inconsistent discussions of externalities. Marshall's externality refers to the impact from activities of other economies, with the typical example the tragedy of the commons. Pigou's externalities refer to the influence of actors on society and the natural environment, such as global warming and intergenerational equity in sustainable development theory. Coase's externality advocates the influence of actors on direct participants, such as the impact of factory sewage on fish farms. The aforementioned studies were also conducted at different levels in terms of extension, which has exacerbated the ambiguity of the concept of externality. Many economists, such as Tibor Scitovsky, Zhang, Yang and others, believe that the definition of externality is a rather vague concept. Hu Shiqing and Wu Jiapei (2011) summed up the research on externalities by Marshall, Pigou, Coase and other scholars and provided two essential features of externalities: that the affected party has “non-participatory decision-making” and “has lacked an effective feedback mechanism.” Based on this, a more comprehensive definition can be given: in a particular economic activity, externalities are the effects on parties that have not participated in the decision and lack an effective feedback mechanism to compensate. The externality is generated and exists in this way.

Despite the unclear definition of externality, current research on the externalities of an industry often does not consider the ambiguity of the concept. Research situations on the topic more common in traditional industries and have provided some inspiring conclusions. However, it is undeniable that traditional industries are relatively simple and their economic activities are relatively stable. For emerging economies which are sharing economies, new resource allocation methods, transaction circulation modes, diversified needs and business types may be impossible under vague framework for comprehensive and accurate analysis. It is too easy to catch only a few details and conduct incomplete study under the current environment.

The ambiguity of the externality concept creates difficulties for quantitative analysis of the presence and origin of externalities. Existing quantitative or empirical studies have focused on the FDI industry and often examine externalities through the spillover effect of technology or knowledge (Aitken and Harrison, 1999; Wen and Liang, 2011). Another type of study is the discussion of the endogenous nature of externalities (Mitchell and Skrzypacz, 2006). The empirical approach of the former is generally to reflect the marginal effects of variables through regression coefficients; the latter is usually derived through mathematical model derivation. However, the aforementioned research methods cannot explain the source of externalities well. First, the basic idea in the regression method is weak in explaining the cause and effect relationship. Second, the complexity of the concept of externalities makes it difficult to abstract a reasonable mathematical expression. Finally, the sharing economy studied in this paper is a new economic form.
At present, there are no mature technical means or models that can be used to describe the external effects of an activity, and it is impossible to explore the sources of externalities. Based on the three aforementioned points, we need to adopt a new method to explore the sources of negative externality in the sharing economy. Qualitative comparative analysis (QCA) provides useful ideas for this study. This paper believes that the QCA method is more suitable for the study than the traditional statistical methods, with the main reason that QCA has a significant advantage in explaining the causality.

At present, research on the sharing economy has focused primarily on the drivers (Shirky, 2009), impact assessments (European Parliament, 2015; Greenwood and Wattal, 2015), business model characteristics (Henten and Windekiilde, 2015), effective management (Zhang, 2016; Chen, 2016; Liu and Wang, 2016) and so on. However, mature studies on externalities of the sharing economy are rare. At present, there are no professional papers or authoritative reports concerning the externalities of the sharing economy, with references to the negative externalities of the sharing economy often found in Web news reports. The reasons for the separation of realistic and theoretical research may include that the sharing economy has appeared recently, and its understanding is still deepening. Also, as aforementioned, externality is a relatively vague concept, and therefore, although we can recognize its existence, it is difficult to analyze with mature methods. The over-innovation of research objects has also made it difficult to find effective research methods. On the other hand, there has been an endless stream of research on the sharing of economic supervision and governance. In fact, these studies have illustrated that the sharing economy has brought a series of benefits to society, while simultaneously producing certain negative consequences. A large part of these consequences can be summarized and interpreted with negative externalities. However, the existing research neglects analysis of problem causes and discusses the governance methods directly, lacking innovative conclusions.

Therefore, in the discussion of the negative externalities of the sharing economy, we first need to establish a normative framework to study the scope of externalities, the production mechanisms and the extension of performance. The aforementioned definitions have not been widely recognized, but inspiration can be drawn from these studies. First, externalities are related to human behavior; Marshall, Pigou and Coase all involved this aspect in the definition. Therefore, the behavior of the economic subject can be used as the starting point of externality study. Second, economic entities can be divided rationally as the generator and the receiver of externality. This division can better determine whether the externality exists. Finally, on the basis of this judgment, to explain the causes or mechanisms of externalities, it is possible to start by distinguishing the behaviors that generate externalities and examining what factors have caused these behaviors.

3. Analysis framework
Positive externalities in the sharing economy can be clearly defined and reflected in the reduction of transaction costs and improvement of transaction efficiency. This is determined by the original intention of the sharing economy. Negative externalities, however, have a variety of expressions, for example, man-made destruction of shared resources (such as bicycles and rooms), indiscriminate use of shared bicycles affecting traffic operation and security problems arising from online car rental. The aforementioned issues have different performances and involve different levels of interpretation. Man-made destruction of shared resources involves the issue of public goods, the illegal parking of shared bicycles involves the issue of property rights separation and the security problem caused by online car rentals is mainly from the assumption of reasonable persons who act as profitable agents.
It can be seen that the sources of negative externalities in the sharing economy are mainly related to participant behaviors. Immoral behavior means that “self-interested individuals are affected by certain factors and violate general principles such as honesty and reliability because the environment allows him to do so with impunity” (Pang, 2004). Therefore, the focus of research on the negative externalities of the sharing economy lies in the analysis of conditions for immoral behaviors. On the other hand, the Mandeville Paradox points out that private evil can be transformed into public interest. This can reveal that some deterrence is not necessarily the source of diseconomy; on the contrary, it is possible that strong self-interest promotes public interest. This kind of thinking is also reflected in the first theorem of welfare economics. That is, if companies pursue interests and every individual pursues his/her own maximization of benefits, market resource allocation can reach Pareto optimality.

Therefore, in the analysis of negative externalities, it is not possible to consider only the subjective behavior of the sharing economy as an independent variable, but other variables must also be included. In addition to the costs and benefits of self-interested behavior, some new issues arising from the sharing economic business model should also be considered, including property rights separation, public goods issues and information asymmetry.

4. Research methods and processes
4.1 New idea: qualitative comparative analysis
4.1.1 The origin and application of qualitative comparative analysis. QCA is an analysis method for small and medium-sized samples of case studies, first proposed in social science research in the 1980s by Charles Ragin. This method is based on case studies and can systematically analyze small data samples. It combines the strengths of both qualitative and quantitative analysis methods, treating each case of the study as a whole and analyzing the interpretation conditions and configurations in the case. In the early use of the method, a clear set qualitative comparative analysis (csQCA) technique was mainly used. However, this technique can only deal with binary variables, that is, the causality condition of the analysis, and the result value must be calibrated to 0 or 1, facilitating the loss of variable information and production of contradictory configurations and increasing the difficulty and challenge of analysis (Cronqvist and Berg-Schlosser, 2009). To overcome this defect, Ragin (2008) proposed the fuzzy set qualitative comparative analysis (fsQCA) technology, adopting a fuzzy set to indicate the degree of occurrence of results and the interpretation conditions. In principle, any value between 0 and 1 can be assigned. The loss of information in the process of data transformation can be avoided, and the actual case situation can be more accurately reflected. Therefore, this paper chose the fuzzy set QCA method.

4.1.2 The principle of qualitative comparative analysis. The core logic of QCA is derived from set theory. If a problem or phenomenon to be studied is viewed as a set, the preconditions causing the phenomenon are subsets. For example, when analyzing the source of negative externalities in the sharing economy, “negative externalities in the sharing economy” is a set, and the subsets of preexisting conditions can be “gains from immoral behavior,” “products or services with the nature of public goods,” “limited industry constraints” and so on. On the technical level, the QCA method uses a Boolean algebra algorithm to find certain affiliation relationships that exist among the collections through a certain number of multi-case comparisons. As there are equivalent multiple paths or solutions in research (Ragin, 2000, 2014; Rihoux and Ragin, 2009; Fiss, 2011), the QCA method may explore multiple relationships between combinations of condition variables.

Compared with the general regression method, the QCA method is more helpful in identifying the multiple concurrency causes of the results. The configuration theory and
model (three interactions) have higher explanatory power than the contingency theory and model (binomial interaction), which has already formed consensus in the academic community (Dess et al., 1997). Traditional contingency methods only analyze the relationship between a certain independent variable and a dependent variable. The QCA method with the configuration perspective can handle the multi-dimensional and multi-causal conditions of interdependence, configuration equivalency, causal asymmetry and other causal complexity issues (Fiss, 2007). Differences between using the general regression method and the QCA method to study “the source of negative externalities in the sharing economy” can be more intuitively represented by Figures 1 and 2, respectively.

From Figures 1 and 2, we can see that the general regression analysis focuses on the marginal net effect of the independent variable on the dependent variable; therefore, the complex causal relationship between independent variables cannot be explained. The QCA method adopts a configuration perspective and reflects how the condition variables (corresponding to the independent variables) cause the result variables (dependent variables) to occur in different combinations.

The calculation principle of QCA is based on Boolean algebra. A variable of 1 indicates that a certain condition occurs, and 0 indicates that a certain condition does not occur. A value of 0 is represented by “¬”. The logical operator “•” means “and,” and “+” means “or.” The symbols “=” or “→” mean “cause.” For example: “A•B + C¬D → Y” indicates that conditions where A and B coexist, or C occurs and D does not occur, may cause Y to occur.

**Figure 1.**
Schematic diagram of the general regression method

**Figure 2.**
Schematic diagram of QCA[1]
Measuring the reliability of QCA results can be performed through two indicators: coverage and consistency, calculated as follows:

\[
\text{Coverage}(X_i \leq Y_i) = \frac{\sum (\min(X_i, Y_i))}{\sum Y_i}
\]  

\[
\text{Consistency}(X_i \leq Y_i) = \frac{\sum (\min(X_i, Y_i))}{\sum X_i}
\]

In the formula, \(X_i\) represents a set of certain condition variable combinations, and \(Y_i\) represents a case result set. The coverage rate examines the extent to which \(X_i\) can form the necessary conditions for \(Y_i\); consistency examines how much \(X_i\) can deduce the outcome of the \(Y_i\) set.

4.1.3 General steps for qualitative comparative analysis. QCA begins with the selection of cases based on research questions. This is a process of constant interaction between theory and experience. As QCA analyzes conditional combinations rather than cases, conditional variables need to be determined according to various strategies and corresponding theories. The purpose is to derive the data combination of condition and result variables then build a truth table to reflect all combinations of condition variables and result variables. In this process, contradictory conditional combinations may be encountered, requiring calibration. Finally, the truth tables are simplified according to Boolean algebra to obtain various combinations of conditions that lead to the occurrence of result variables.

4.2 Case selection
Case selection is the basis for OCA. As the sharing economy is a new kind of economic model, internet news reports can fully reflect its characteristics, and research results may be in touch with reality. This article searched keyword “sharing economy” to extract typical negative cases of the sharing economy that have occurred in the past two years. To ensure research quality, cases were screened in two rounds. In the first round, cases that met the following conditions were selected: cases listed on mainstream media and cases described in detail to ensure that relevant information could be obtained. In the first round, 39 cases were selected. In the second round of screening, cases with the same event attributes were removed. For example, only one case of privatization of shared bicycles was retained. Finally, 15 cases were coded (refer Table I for details).

4.3 Truth table construction
The construction of a truth table is a key step in QCA. Based on the analysis framework proposed in the previous section, this article focuses on the condition variables that result from negative externalities in the shared economy. This includes the loss of gains, expected compensation, convenience of defeat and degree of public goods. There are six aspects to the degree of separation of property rights and the restraint mechanism of failure. The outcome variable chosen is the severity of negative externalities. Based on the aforementioned variables, a truth table was formed. The specific assignment method is as follows.

4.3.1 The severity of negative externalities (RI). The severity of negative externalities was the outcome variable, using a three-value assignment scheme with 0 as the lowest severity, 0.5 moderately severe and 1 the highest. Assignment rules were based on the actual damage caused in the case. Inflicting harm to others, causing serious loss of property or spirituality to others, or repeatedly behaving in a highly unpromising manner resulting in strong social adverse effects, etc., were assigned a value of 1. The variable was set to 0 for scenarios that did not directly harm others. The remaining cases were assigned 0.5. For
example, Case 9 mentioned “Didi’s driver escaping after hitting someone” causing harm to the life of others and was assigned the value of 1. Case 7: sharing bicycles near Beijing East Station piled up into hills, a group of irregular behaviors, was assign 0. It is noteworthy that the severity of negative externalities in some cases did not have a distinct degree of differentiation, so the authors discussed in detail during the assignment process until a consensus was reached. Subsequent variables that involved segmented assignments follow the same approach.

4.3.2 Immoral behavioral benefits (Benefit – C1). Immoral behavioral benefits are an important part of the traditional analysis perspective of cost–benefit analysis. The assignment is more complicated because some benefits are material gains, while some are psychological satisfaction. To better characterize these differences, a four-value valuation scheme was used for the return factor, which was assigned to 0, 0.33, 0.67 and 1 based on low or high returns. The assignment rules were divided according to the physical or psychological gains of the immoral subject, for example, the behavior of obtaining economic benefits directly by replacing the shared bicycle’s QR code was assigned a value of 1 (Case 15).

4.3.3 Expected compensation (Compensation – C2). Immoral behavioral benefit and expected compensation serve as important components of traditional analytical perspectives. Expected compensation mainly reflects the cost aspect. According to the expected level of compensation, values were assigned 0, 0.5 or 1, from low to high. The assignment rules referred to the value of damage to physical objects and the behavioral characteristics of the participants. For example, Didi’s driver made a collision and escaped the scene (Case 9); “escape” was used to determine that the driver’s expected

<table>
<thead>
<tr>
<th>Case no.</th>
<th>Case title</th>
<th>Source website</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Exposure to a student of Shanghai Theater Academy who destroyed my entire home with Airbnb</td>
<td>Sohu Tourism</td>
</tr>
<tr>
<td>2</td>
<td>Master stole bike and sprayed the bike into black</td>
<td>Sina Shanghai</td>
</tr>
<tr>
<td>3</td>
<td>The two guys throw the ofo sharing bike into the river and broadcast it</td>
<td>NetEase Digital</td>
</tr>
<tr>
<td>4</td>
<td>Because the sharing of bicycles affected the business, the electric tricycle driver burned the bicycle and was detained</td>
<td>Sina News Center</td>
</tr>
<tr>
<td>5</td>
<td>Needles appear on shared bicycle seats!</td>
<td>Beijing Time</td>
</tr>
<tr>
<td>6</td>
<td>Nanning: villages in the city where the bicycles are trapped in and changed to children’s toys</td>
<td>Guangxi news site</td>
</tr>
<tr>
<td>7</td>
<td>Sharing bicycles near Beijing East Station piled up into hills</td>
<td>Beijing Youth Daily (electronic version)</td>
</tr>
<tr>
<td>8</td>
<td>A woman was called to the crematorium by Didi’s driver, every day for half a month</td>
<td>Sohu.com</td>
</tr>
<tr>
<td>9</td>
<td>Traffic control big data disclosure about online car rental’s security short board, Didi mode is a safety “dead end”</td>
<td>Oriental headlines</td>
</tr>
<tr>
<td>10</td>
<td>Didi’s driver harassed female clients and have been detained by the Public Security Bureau</td>
<td>cc.cn</td>
</tr>
<tr>
<td>11</td>
<td>Driver of Didi arrives late and causes passengers to miss flights. Who will pay ¥2,400?</td>
<td>Tencent</td>
</tr>
<tr>
<td>12</td>
<td>Didi driver fascinated into a female passenger and claimed to use an Apple 6S to change her boyfriend to him</td>
<td>Tencent</td>
</tr>
<tr>
<td>13</td>
<td>Didi booking design scam, and old driver was also cheated ¥680</td>
<td>NetEase Finance</td>
</tr>
<tr>
<td>14</td>
<td>Shared bicycle was posted “Stolen Money QR Code” 360 exposure three scan code scam</td>
<td>Sina Finance</td>
</tr>
<tr>
<td>15</td>
<td>Retirement of the shared bicycle deposit was plunged and many people have been deceived! The cheated people have done this . . .</td>
<td>Tencent Weibo</td>
</tr>
</tbody>
</table>
compensation for the event was high, so the value was assigned 1. In Case 8, a woman was called to the crematorium by the driver every day for half a month only because of bad feedback to the driver. The behavior in this case exhibits a continuous process, indicating that the participant’s expected compensation for the matter was extremely low, so the value was set to 0. Other cases, such as the destruction of shared bicycles or housing, were set to 0.5.

4.3.4 Convenience of immoral behavior (Convenience – C3). The foregoing reveals that the sources of negative externalities are largely related to behavior. Therefore, the convenience of immoral behavior reflects the possibility of negative externalities. According to the degree of convenience, values of 0, 0.5 and 1 were assigned, from low to high. For example, parking shared bicycles randomly as a habit is highly convenient, and the assigned value was 1; it is less convenient to throw a shared bicycle into the river or burn a shared bicycle, so the assigned value was 0.

4.3.5 Nature of public goods (Public – C4). Characteristics of public goods are a prominent aspect of products and services in the sharing economy. Public goods refer to goods and services provided by the public sector to meet the public needs of society. Public goods have the characteristics of indivisibility, non-competitiveness and non-excludability. Obviously, these characteristics in sharing bicycles is significantly higher than cars, such as rental services represented by Didi, and homes, such as rental services represented by Airbnb. Therefore, cases involving a shared bicycle were assigned a value of 1, and the remaining cases were assigned a value 0.

4.3.6 Separation of property rights (Use or possess – C5). The separation of property rights is an important feature of the sharing economy. The method used to measure the separation of property rights in this paper was whether in each case the participating entity had the right to use the product or service. In cases when there was only the right to use, we believe that the degree of separation of property rights is relatively high. Cases with this condition were assigned to a value of 1 and the rest of the cases 0.

4.3.7 Constraint mechanism (Constrains – C6). To restrict the negative externalities that arise from sharing of products or services and reduce the unnecessary costs to enterprises or society, entities in the sharing economy have established a certain constraint

<table>
<thead>
<tr>
<th>Case no.</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
<th>R1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.33</td>
<td>0.5</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>0.67</td>
<td>0.5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>3</td>
<td>0.67</td>
<td>0.5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0.67</td>
<td>0.5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0.67</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>7</td>
<td>0.33</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0.33</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>0.67</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0.5</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>12</td>
<td>0.33</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table II. Truth table

Sharing economy
mechanism. The two most common mechanisms are the deposit mode and the evaluation mode. This article sets the deposit mode to 1 and the comment mode to 0. From this, we obtained the truth table for this article, shown in Table II.

5. Result analysis

5.1 Single factor discussion

First, we examined the relationship between single conditional variables and negative externalities. The results are shown in Table III. None of the six conditional variables satisfied the conditions shown in equation (1) from the point of view of consistency. From the perspective of coverage, only the variable of expected compensation met the conditions of equation (2). Therefore, a single conditional variable cannot explain the formation of negative externalities in the sharing economy. The results in Table III also indirectly illustrate the necessity of QCA.

5.2 Traditional perspective discussion

Of the six conditional variables selected in this paper, the traditional perspective of cost–benefit analysis identified three which form negative externalities—imoral behavioral benefits, expected costs and convenience of immoral behavior. The purpose of the traditional QCA method is to identify the role of sharing economic characteristics in the negative externalities. Using fs/QCA2.0 software, case codes with continuity of less than 0.8 were deleted, and standard analysis was selected. The output results are shown in Table IV.

In Table IV, coverage (or raw coverage) represents the effect of each factor in each condition combination on the results and is generally not analyzed. The value of unique coverage indicates which combinations are more able to achieve the result, and consistency reflects the number of cases that the condition combination can explain. On the whole, the consistency was higher, exceeding 0.97, which meets the requirements. The overall coverage was low, which may be constrained by the sample size.

Table IV shows two paths of negative externality formation in the sharing economy from the traditional perspective:

<table>
<thead>
<tr>
<th>Conditional variable</th>
<th>Consistency</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benefit</td>
<td>0.620909</td>
<td>0.890482</td>
</tr>
<tr>
<td>Compensation</td>
<td>0.318182</td>
<td>1.000000</td>
</tr>
<tr>
<td>Convenience</td>
<td>0.409091</td>
<td>0.642857</td>
</tr>
<tr>
<td>Public</td>
<td>0.545455</td>
<td>0.750000</td>
</tr>
<tr>
<td>Use or possess</td>
<td>0.454545</td>
<td>0.625000</td>
</tr>
<tr>
<td>Constraint</td>
<td>0.545455</td>
<td>0.750000</td>
</tr>
</tbody>
</table>

| Solution coverage     | 0.666364    | 0.977333 |
| Solution consistency  | 0.000000    | 1.000000 |

Table III.

Single factor analysis results

<table>
<thead>
<tr>
<th>Conditional variable</th>
<th>Consistency</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benefit</td>
<td>0.620909</td>
<td>0.890482</td>
</tr>
<tr>
<td>Compensation</td>
<td>0.318182</td>
<td>1.000000</td>
</tr>
<tr>
<td>Convenience</td>
<td>0.409091</td>
<td>0.642857</td>
</tr>
<tr>
<td>Public</td>
<td>0.545455</td>
<td>0.750000</td>
</tr>
<tr>
<td>Use or possess</td>
<td>0.454545</td>
<td>0.625000</td>
</tr>
<tr>
<td>Constraint</td>
<td>0.545455</td>
<td>0.750000</td>
</tr>
</tbody>
</table>

Table IV.

Negative externalities of the sharing economy from the traditional perspective

<table>
<thead>
<tr>
<th>Solution</th>
<th>Combination of conditions</th>
<th>Raw coverage</th>
<th>Unique coverage</th>
<th>Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intermediate solution</td>
<td>benefit<em>~compensation</em>~convenience</td>
<td>0.484545</td>
<td>0.484545</td>
<td>0.969091</td>
</tr>
<tr>
<td>(Complex solution)</td>
<td>~benefit<em>compensation</em>convenience</td>
<td>0.181818</td>
<td>0.181818</td>
<td>1.000000</td>
</tr>
<tr>
<td>Solution coverage</td>
<td></td>
<td>0.666364</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solution consistency</td>
<td></td>
<td>0.977333</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Traditional Path 1: Source of negative externalities in the sharing economy = (high immoral behavioral benefits) * (low expected compensation) * (inconvenience of immoral behavior); and

Traditional Path 2: Source of negative externalities in the sharing economy = (low immoral behavioral benefits) * (high expected compensation) * (convenience of immoral behavior).

The two aforementioned paths are in line with experience. That is to say, when immoral behavior has high benefits and low cost, negative externalities will still be generated, even with lack of convenience. On the other hand, if it is convenient to exercise the immoral behavior, then even if the gap between the benefits and costs is not particularly large, negative externalities are also likely to arise. From the size of the raw coverage, it can be seen that the Traditional Path 1 is more likely to occur. However, the aforementioned results do not portray the characteristics of the sharing economy. The question of whether a special model such as the sharing economy has exacerbated negative externalities cannot be explained well. Therefore, the path to negative externalities in the sharing economy should be studied further.

5.3 Comprehensive discussion

In the same way, the aforementioned six conditional variables were put into the same analysis framework. The output is shown in Table V.

The solution consistency in Table V is greater than 0.9, falling within the acceptable range. Table V gives three paths to negative externalities in the sharing economy:

1) Path 1: Source of negative externalities in the sharing economy = (low immoral behavioral benefits) * (low expected compensation) * (convenience of immoral behavior) * (low level of public goods) * (low level of separation of property rights) * (evaluation constraints);

2) Path 2: Source of negative externalities in the sharing economy = (low immoral behavioral benefits) * (high expected compensation) * (convenience of immoral behavior) * (low level of public goods) * (high level of separation of property rights) * (evaluation constraints); and

3) Path 3: (high immoral behavioral benefits) * (low expected compensation) * (inconvenience of immoral behavior) * (high level of public goods) * (high level of separation of property rights) * (deposit constraints).

Of the aforementioned three paths, Paths 2 and 3 complement the traditional path under the sharing economy environment. Among them, Path 2 shows that when there is high convenience in immoral behavior, higher degrees of separation of property rights and impacts of evaluation constraints and lower levels of public goods will reduce negative externalities. On the other hand, Path 3 reveals another economic model that promotes negative externalities: high level of public nature of goods, high level of separation of property rights and the deposit constraint mechanism, similar to shared bicycles.

In addition, Paths 1 and 3 illustrate that the negative externality of the sharing economy is related to specific business models. Specifically, Path 1 illustrates that negative externalities in the sharing economy can result from the following circumstances: a low level of public goods, property rights that are not completely separated and the evaluation constraint mechanism. These characteristics are in line with the characteristics of online car rentals and other short-term rental services. In this kind of environment, the levels of
<table>
<thead>
<tr>
<th>Solutions</th>
<th>Combination of conditions</th>
<th>Raw coverage</th>
<th>Unique coverage</th>
<th>Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intermediate solution</td>
<td>~benefit * ~compensation * convenience * ~public * ~useorpossess * ~constraint</td>
<td>0.120009</td>
<td>0.120009</td>
<td>0.886667</td>
</tr>
<tr>
<td>(Complex solution)</td>
<td>~benefit * compensation * convenience * ~public * useorpossess * ~constraint</td>
<td>0.181818</td>
<td>0.181818</td>
<td>1.000000</td>
</tr>
<tr>
<td></td>
<td>benefit * ~compensation * ~convenience * ~public * useorpossess * constraint</td>
<td>0.272727</td>
<td>0.272727</td>
<td>0.946372</td>
</tr>
<tr>
<td>Solution coverage</td>
<td></td>
<td>0.575455</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solution consistency</td>
<td></td>
<td>0.949025</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table V. Formation path of negative externalities in the sharing economy.
benefits and costs of immoral behaviors have the same effect on negative externalities. The lower the two, the more likely negative externalities will arise. In this environment, the Traditional Path 1 is more likely to form negative externalities. Comparing Path 1 and Path 3 can further determine that increases in the degree of sharing of products or services (high level of public goods and high level of separation of property rights) and more stringent constraint models (deposit constraint) contribute to negative externalities in the sharing economy. This actually forms a kind of paradox. From an empirical point of view, the higher the degree of sharing, the more stringent the constraint mechanisms are generally required to be. However, research in this paper shows that it is precisely this kind of strict constraint mechanism that causes higher degrees of sharing to form negative externality.

6. Recommendations
The aforementioned conclusions indicate that the common business models in the current sharing economy cannot spontaneously avoid negative externalities but instead have the potential to promote negative externalities. Therefore, the corresponding governance model is needed to ensure a sound development of the sharing economy. This paper proposes the following recommendations.

First, highlight the platform corporate responsibility. The sharing economy makes economic activities not only show the characteristics of happening at any time and everywhere. Moreover, the patterns, groups and fields in which they occur are also expanding. In this case, it is impossible to achieve effective governance of the sharing economy through a top-down perfect system. Therefore, it is necessary for the platform enterprises to assume certain social responsibilities and supervise the main participants of the sharing economy. The key is to determine the degree of shared services by the platform enterprises to adapt to the corresponding constraint mechanism.

Second, establish a multi-subject governance framework. The sharing economy is a platform for sharing, co-constructing and co-governing in which the whole society participates. Therefore, in addition to the government and enterprises, there should also be public participation. Therefore, it is necessary to establish a multi-governance system including the government, enterprises and public. The multi-governance system needs to clarify the governance boundaries, scientific division of labor and efficient cooperation among the various entities. Guiding the public to participate in governance actively, making it not only a participant in the sharing economy but also a manager of the sharing economy.

Finally, build a multi-level safeguard and constraint mechanism. This part includes two aspects, one is to establish and improve industry norms and laws; the other is to use new technologies such as big data and cloud computing to incorporate individual behavior into the personal credit system and form a new constraint mechanism.

7. Conclusion
The rapid development of the sharing economy has improved the efficiency of resource utilization, contributing to the quality of economic development. However, various undeniable problems have arisen in the operation and development of the sharing economy at home and abroad, with very prominent negative externalities brought by its disorderly development.

Therefore, this article analyzed the source of negative externalities in the sharing economy. Through the QCA method, three sources of negative externalities were obtained. We believe that sharing of products or services and restraint mechanisms work together to create conditions for negative externalities in the sharing economy. Results found that sharing economy modes with less sharing and more relaxed constraint mechanisms, such as online vehicles and network short-term rentals or modes with higher sharing nature and
strict restraint mechanisms, such as shared bikes, both promoted negative externality. Finally, corresponding policy recommendations are proposed for this issue. This paper started with practical problems and proposed how various factors interact and ultimately contribute to the formation of negative externalities in the sharing economy. The conclusion of this paper provides a new path for the healthy development of the sharing economy and avoids its adverse effects. It also provides a theoretical basis for the government governance of the sharing economy.

Note
1. Figure 2 only shows the result form of the QCA method and does not represent the final results of this study.

References


Zhang, X. (2016), Research of China Systemically Important Banks Regulation of Negative Externalities, Shanxi University of Finance and Economics.


Further reading


Pigou, A.C. (1999), The Economics of Welfare, China Social Sciences Pu.


Corresponding author
Wenjun Jing can be contacted at: jwj881216@sina.com
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

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
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:

\[
\begin{cases} 
C_1(x), x_0 \leq x \leq x_1 \\
\vdots \\
C_n(x), x_{n-1} \leq x \leq x_n 
\end{cases}
\]

(1)

Where each \( C_i \) is a cubic function as:

\[
C_i = a_i + b_i x + c_i x^2 + d_i x^3
\]

(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:

\[
\begin{cases} 
a_i = y_i \\
c_i = \frac{1}{2} S''_i(x_i) = m_i \\
b_i = \frac{y_{i+1} - y_i}{h_i} - \frac{h_i}{2} m_i - \frac{h_i}{6} (m_{i+1} - m_i) \\
d_i = \frac{m_{i+1} - m_i}{6h_i}
\end{cases}
\]

(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:

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

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

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

3: calculate the Euclclidean distance $d(x', o')$ of $x'_{5 \times 1}$ with all the other data samples $o'_{5 \times 1}$
```
4: locate the K-th nearest point \( N_k(x') \) of \( x'_{5 \times 1} \) and set the distance between the two points as \( k-dist(x', o) \) of \( x'_{5 \times 1} \)

5: calculate the reachable distance \( reach-dist(x', o') \) between \( x'_{5 \times 1} \) and points \( o' \) based on equation: \( reach-dist(x', o') = \max(k-dist(x'), d(x', o')) \)

6: calculate the local outlier density \( lrd(x') \) based on equation

\[
lrd(x') = \frac{1}{|N_k(x')|} \sum_{o \in N_k(x')} \frac{lrd(o')}{lrd(x')}
\]

7: calculate the LOF \( lof(x') \) based on equation \( lrd(x') = \frac{1}{|N_k(x')|} \sum_{o \in N_k(x')} \frac{lrd(o')}{lrd(x')} \)

8: if \( lof(x') > threshold \) then

9: \( D(x') \leftarrow \text{anomaly} \)

10: end if

11: end for

12: 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 practitioners 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 lower 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:
\[ \text{supp}(x_1) = \frac{|\{t \in T; x \in t\}|}{|T|} \quad (4) \]

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 \( x \). 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 \( \emptyset \), 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:

\[ \text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} \quad (5) \]

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 drawbacks 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](image.png)

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 categories 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 lightweight version of the famous metrics named Information Gain. The Gini impurity of a subset can be calculated as:

\[ I(p) = \sum_{i=1}^{I} p_i \sum_{k \neq i} p_k = \sum_{i=1}^{I} p_i(1 - p_i) = 1 - \sum_{i=1}^{I} p_i^2 \]  

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 uses 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 \( v = (x_1^T, x_2^T, \ldots, x_n^T) \), 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:

\[
y_i = \theta_0 + \theta_1 x_{i1} + \ldots + \theta_m x_{im} + \epsilon_i
\]

where \( \epsilon_i \) 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:

\[
Y = h_\theta(X) = \theta^T X + \epsilon
\]

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

\[
J(\theta^T) = \min (\sum_{i=1}^{m} (h_\theta(x_n) - y_i)^2)
\]

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

\[
\theta = (X^T X)^{-1} X^T Y
\]

In this way, the values within matrix \( \theta \) 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).
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...
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.
- 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.
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\} \rightarrow \{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\} \rightarrow \{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\} \rightarrow \{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:

- there are more female population who are at school than that of male population;
- the female population are more likely to be married than male population; and
- 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


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

For instructions on how to order reprints of this article, please visit our website: www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: permissions@emeraldinsight.com.
International Journal of Crowd Science

Number 2

85 Editorial board

86 Application of biclustering algorithm to extract rules from labeled data
Zhang Yanjie and Sun Hongbo

99 Quality-time-complexity universal intelligence measurement
Jing Liu, Zhiwen Pan, Jingce Xu, Bing Liang, Yiqiang Chen and Wen Ji

108 Missing observation approximation for spatio-temporal profile reconstruction in participatory sensor networks
Assad Mehmood, Kashif Zia, Arshad Muhammad and Dinesh Kumar Saini

123 Towards data-driven software engineering skills assessment
Jun Lin, Han Yu, Zhengxiang Pan, Zhiqi Shen and Lizhen Cui

136 A scientometric analysis of e-participation research
Tuotuo Qi, Tianmei Wang, Yanlin Ma, Wei Zhang and Yanchun Zhu

149 Negative externalities in the sharing economy: sources, paths and recommendations
Wenjun Jing and Baowen Sun

164 Anomaly data management and big data analytics: an application on disability datasets
Zhiwen Pan, Wen Ji, Yiqiang Chen, Lianjun Dai and Jun Zhang