Linear Correlation Analysis of Numeric Attributes for Government Data

Ying Chen¹, Guochang Gu¹, Tianyang Lv¹, Shaobin Huang¹, Jun Ni¹,²
¹School of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China
²Department of Computer Science, the University of Iowa
{guguochang, lvitianyang, huangshaobin, nijun}@hrbeu.edu.cn

Abstract
To analyze the linear correlations of numeric attributes of government data, this paper proposes a method based on the clustering algorithm. A clustering method is adopted to prune outliers and the linear correlation analysis is performed for each cluster, instead for the whole dataset. In this way, the method can obtain multiple correlations between the same two attributes. The paper presents the experiment on the government social security data. Experimental results show that the proposed method is much better than the traditional regression analysis and association rule analysis.

1. Introduction

The correlation analysis of large datasets has been widely adopted in the commercial field [1]. A typical example is the discovery of the “beer and diaper” rule. However, little effort has been spent on the analysis of the correlations that lie in government data, although this topic has important application value. For instance, Chinese auditors usually handle the large government databases without meta-data. In that case, the correlation analysis of attributes can help them understand datasets and discover illegal actions.

Compared with the commercial data, government data shows its characteristics in the following aspects: its enormity, over G bytes in most cases; most attributes are numeric and multiple linear correlations between numeric attributes.

A traditional method of correlation analysis is the linear regression method [2]. The regression analysis technique originates from the Gaussian Least square method. It analyzes the relationship between the independent variable and the attributive variable. The technique estimates the average of the attributive variable according to the known independent variable. That is, it explains the changes of one variable through the changes of one or more other variables. One variable linear regressive is the simplest form, which only considers the case with just one independent variable and one attributive variable. However, it is influenced greatly by outliers.

In the field of data mining, the association rule analysis technique is adopted to discover the relationships among attributes [3] and it performs quite well for the categorical or basket attribute. To apply this technique to the numeric attribute, the numeric attribute must be converted to the categorical attribute that corresponds to different sections [4]. Therefore, the discovered rules show the correlation among these sections, and cannot reflect the accurate relationship between the numeric attributes.

To overcome the drawbacks of regression analysis and association rule analysis, this paper proposes a new method for discovering the linear correlations between numeric attributes in the government database. The basic idea of the proposed method is the application of the cluster analysis technique on the government dataset. As an unsupervised technique, the cluster analysis can find the congregation of data without prior knowledge. Small clusters can be pruned as outliers [5], while the remaining clusters reflect different distribution patterns. Therefore, we can discover the attribute correlations from each resulting cluster. Thus we can handle some difficult cases, like when more than one correlation exists between attributes and if the correlation exists in just a small part of the dataset.

Thus, the proposed method applies the X-means algorithm [6] in analyzing the sampled government dataset; then discovers the related attributes from each resulting cluster; and determines the linear correlation between the related attributes.

The rest of the paper is organized as follows. After introducing related works in Section 2, Section 3 introduces the proposed linear correlation analysis method. Then, Section 4 outlines an experiment on the government data and Section 5 concludes the paper.
2. Related works

The linear correlation analysis is a method used to analyze the linear relationship of two variables and their degree of relationship. The linear correlation of variables \( X \) and \( Y \) is described by the correlation coefficient \( r \), and \( r \) is computed through Equation (1):

\[
    r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}} = \frac{l_{XY}}{\sqrt{l_{XX}l_{YY}}}
\]

(Figure 1 shows the different cases of linear correlation. In Figure 1.(a), data distributes in an ellipse while the \( X \), \( Y \) values change in the same direction. This is a typical case corresponding to the positive correlation with \( 0 < r < 1 \). Figure 1.(b) corresponds to the negative correlation with \( -1 < r < 0 \), Figure 1.(c) corresponds to the perfect positive correlation with \( r = 1 \), and Figure1.(d) corresponds to the perfect negative correlation with \( r = -1 \).)

Thus, we can decide whether variables \( X \) and \( Y \) show linear correlation by evaluating the value of \( r \). If \( | r | > \theta \), where \( \theta \) is a threshold, the linear correlation exists.

However, outliers can greatly reduce the performance of the correlation analysis. Since the value of \( r \) is influenced by these outliers that are extremely far away from most of the others, this results in the mistakes in the discovered correlations.

In recent years, the clustering analysis technique has experienced rapid improvements in many aspects. For instance, an \( X \)-means algorithm is proposed to overcome the \( K \)-means’ shortcomings in its high dependence on pre-deciding the appropriate number of resulting clusters \( k \). Due to its efficiency and effectiveness, the paper will adopt an \( X \)-means algorithm for analyzing the government data. Also, some other methods, like CURE [5] and DBScan [7], prune the small clusters or the clusters could grow slowly as outliers.

3. The linear correlation analysis of numeric attributes

This paper adopts the \( X \)-means algorithm to pre-handle the government data. As an improvement of the \( K \)-means algorithm, an \( X \)-means algorithm requires, but is not restricted to, the pre-decided number of clusters \( k \). It will split the center of a selected cluster into two children branches and decide whether a child branch should survive. In this way, a more appropriate number of clusters can be automatically decided.

Thus, we can decide whether variables \( X \) and \( Y \) show linear correlation by evaluating the value of \( r \). If \( | r | > \theta \), where \( \theta \) is a threshold, the linear correlation exists.

However, outliers can greatly reduce the performance of the correlation analysis. Since the value of \( r \) is influenced by these outliers that are extremely far away from most of the others, this results in the mistakes in the discovered correlations.

In recent years, the clustering analysis technique has experienced rapid improvements in many aspects. For instance, an \( X \)-means algorithm is proposed to overcome the \( K \)-means’ shortcomings in its high dependence on pre-deciding the appropriate number of resulting clusters \( k \). Due to its efficiency and effectiveness, the paper will adopt an \( X \)-means algorithm for analyzing the government data. Also, some other methods, like CURE [5] and DBScan [7], prune the small clusters or the clusters could grow slowly as outliers.

3. The linear correlation analysis of numeric attributes

This paper adopts the \( X \)-means algorithm to pre-handle the government data. As an improvement of the \( K \)-means algorithm, an \( X \)-means algorithm requires, but is not restricted to, the pre-decided number of clusters \( k \). It will split the center of a selected cluster into two children branches and decide whether a child branch should survive. In this way, a more appropriate number of clusters can be automatically decided.

The clustering result of the \( X \)-means algorithm is refined through the following process. All small clusters whose size is less than 5% of the data’s total number are pruned as outliers. This will remove the disturbance coming from the outliers before correlation analysis.

Since different clusters reflect different distribution situations, the same pair of attributes may show different correlations in these clusters. Therefore, we perform a linear correlation analysis for every cluster instead of analyzing the whole dataset. In this way, we can obtain multiple correlations between the same two attributes.

In analyzing the linear correlations of a cluster \( C_i \), the correlation coefficient \( r \) is computed for all the pairs of numeric attributes, and the analysis names the pairs whose \( r \) values are larger than the correlation threshold \( \theta \) as the candidate correlation.

Then, two ways can be adopted to decide the linear function \( f \) that best reflects the candidate correlation. One is the regression analysis, while the other is analyzing the data that is the nearest and afar that is the nearest and the most far away from \( C_i \)’s center. If the ratio of data belonging to \( C_i \) that satisfies the function \( f \) is larger than the support threshold, \( f \) is rejected.

So the process of linear correlation analysis based on the \( X \)-means algorithm is stated in Figure 2, where Step 2 is the clustering analysis, Step 3 is the refining step and Steps 4 through 8 represent the linear-correlation analyzing processes of all clusters.
1. Determine the value of the correlation threshold \( \theta \) and the support threshold \( \varepsilon \);
2. Apply the \( \times \)-means algorithm to analyze \( N \) sampled data and \( k \) clusters \( \{C_1, C_2, \ldots, C_k\} \) are obtained;
3. For any cluster \( C_i \), if its size is \( n_i < N * 5\% \), prune \( C_i \) as outliers, and \( k = k - 1 \);
4. For a remaining cluster \( C_i \), compute the correlation coefficient \( r \) of any pair of its numeric attributes, and add attributes \{\( col_p \), \( col_q \)\} into \( L_{corre} \) if \( r(col_p, col_q) > \theta \);
5. If \( |L_{corre}| > 1 \), go to Step 6, otherwise go to Step 4;
6. Compute the linear correlation function \( f(col_p, col_q) \), and add \( f(col_p, col_q) \) to \( F_{corre} \) and go to Step 7;
7. Examine each data of \( C_i \) to verify all function \( f(col_p, col_q) \) of \( F_{corre} \) and output all \( f(col_p, col_q) \) whose support degree is larger than \( \varepsilon \);
8. Go to Step 4 if there exists a cluster that has not yet been processed.

**Fig. 2** Process of linear correlation analyzing

### 4. Experiment and Analysis

In the Experiment, the paper adopts a table, termed AB07 from the social security database, which follows the LB101 standard of the Chinese government. The Table AB07 shows the social security payment status of each enterprise and its employees every month. These numeric attributes of the Table and their descriptions are stated in Table 1. Then, we sample the data from the database of this table, and obtain 19469 records from over 200,000 records.

#### Table 1. The numeric attributes of AB07 and their descriptions

<table>
<thead>
<tr>
<th>Column</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>col 3</td>
<td>Number of employees of an enterprise</td>
</tr>
<tr>
<td>col 4</td>
<td>The base rate of employee payment</td>
</tr>
<tr>
<td>col 5</td>
<td>The base rate of enterprise payment</td>
</tr>
<tr>
<td>col 6</td>
<td>Employee’s due payment each month</td>
</tr>
<tr>
<td>col 7</td>
<td>Enterprise’s due payment for its employee each month</td>
</tr>
<tr>
<td>col 8</td>
<td>Employee’s remedy payment for government each month</td>
</tr>
<tr>
<td>col 9</td>
<td>Enterprise’s remedy payment for its employees</td>
</tr>
<tr>
<td>col 10</td>
<td>Refundment of Employee’s payment</td>
</tr>
<tr>
<td>col 11</td>
<td>Refundment of Employee’s payment for government</td>
</tr>
<tr>
<td>col 12</td>
<td>Refundment of Enterprise’s payment for government</td>
</tr>
</tbody>
</table>

Figure 3 is the visualization result of the attributes in \( col 9 \) and \( col 11 \) of the sampled data. It can be seen that some extremely distant outliers exist. As stated in the former part, the value of \( r \) will be greatly influenced by these outliers, which results in the mistakes of the discovered correlations.

#### Table 2. Clustering result of sampled data

<table>
<thead>
<tr>
<th>C_0</th>
<th>C_1</th>
<th>C_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n_i )</td>
<td>18608</td>
<td>361</td>
</tr>
</tbody>
</table>

Then, we adopted two ways to perform a linear analysis on the cluster \( C_0 \), one is the regression analysis, termed the Regression Analysis of \( C_0 \). The other method is based on data near and far from \( C_0 \), and is termed the Correlation Analysis of \( C_0 \). To discover the significant correlations, we set \( \theta = 0.96 \) and \( \varepsilon = 0.99 \).

We perform the traditional regression analysis on the whole set of the sampled data, termed the Tradition Regression Analysis of \( C_0 \).

Table 3 is a result comparison among these three ways. It can be seen that the Correlation Analysis of \( C_0 \) is the best player for all the discovered correlations. For instance, the correlation \( ”col6=0.08*col4“ \) means the employee’s due payment each month equals 8% of their base; \( col7=0.03*col4 \) means enterprise’s due
payment for its employee equals 3\% of the base, while the enterprise base equals the employees' base. We also analyze the association rule of the sampled data. However, as analyzed, the result is quite unacceptable. For example, the correlation \( \text{col}_5 = 1.0 \ast \text{col}_4 \) is stated as the different rules, like \( \text{col}_5[-\infty, 2000] \rightarrow \text{col}_4[-\infty, 2000] \) and \( \text{col}_5[2000, 15000] \rightarrow \text{col}_4[2000, 15000] \).

**Table 3. Comparison of linear correlation analysis**

<table>
<thead>
<tr>
<th>Traditional Regression Analysis</th>
<th>Regression Analysis of Cluster ( C_0 )</th>
<th>Correlation Analysis of ( C_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{col}_5 = 1.0 \ast \text{col}_4 )</td>
<td>( \text{col}_5 = 1.0 \ast \text{col}_4 )</td>
<td>( \text{col}_5 = 1.0 \ast \text{col}_4 )</td>
</tr>
<tr>
<td>( \text{col}_6 = 0.08 \ast \text{col}_4 - 10.61 )</td>
<td>( \text{col}_6 = 0.08 \ast \text{col}_4 + 0.02 )</td>
<td>( \text{col}_6 = 0.08 \ast \text{col}_4 )</td>
</tr>
<tr>
<td>( \text{col}_7 = 0.0089 \ast \text{col}_4 + 427.07 )</td>
<td>( \text{col}_7 = 0.03 \ast \text{col}_4 + 3.62 )</td>
<td>( \text{col}_7 = 0.03 \ast \text{col}_4 )</td>
</tr>
<tr>
<td>( \text{col}_8 = 0.1408 \ast \text{col}_6 + 956.37 )</td>
<td>( \text{col}_8 = 0.19 \ast \text{col}_4 - 16.23 )</td>
<td>( \text{col}_8 = 0.19 \ast \text{col}_4 )</td>
</tr>
<tr>
<td>( \text{col}_9 = 0.1117 \ast \text{col}_6 + 428.24 )</td>
<td>( \text{col}_9 = 0.375 \ast \text{col}_6 + 3.61 )</td>
<td>( \text{col}_9 = 0.375 \ast \text{col}_6 )</td>
</tr>
<tr>
<td>( \text{col}_10 = 6.3215 \ast \text{col}_7 + 258.61 )</td>
<td>( \text{col}<em>{10} = 6.3266 \ast \text{col}</em>{7} - 36.02 )</td>
<td>( \text{col}<em>{10} = 6.3333 \ast \text{col}</em>{7} )</td>
</tr>
</tbody>
</table>

**5. Conclusions**

To analyze the linear correlations of the numeric attributes of government data, this paper proposes a linear correlation analysis method based on a clustering algorithm. And the experimental results show that the proposed method is much better than either the traditional regression analysis or the association rule analysis.

Future works lie in the study of correlation analysis for the attributes with multiple types.

**References**

[1] Jiawei Han, etc., *Data mining: concepts and techniques*, second edition. Elsevier Inc., 2006.

[2] Linear regression


