Finding dependencies in the corporate environment using data mining

Anastasia Kozlova*, Vladislav Kukartsev1,2, Vladimir Melnikov3, Georgiy Kovalev1, and Alexander Stashkevich1

1 Reshetnev Siberian State University of Science and Technology, 660037 Krasnoyarsk, Russia
2 Bauman Moscow State Technical University, 105005 Moscow, Russia
3 ITMO University, 197101 St. Petersburg, Russia

Abstract. The article analyses the influence of factors of the work environment, as well as the non-work environment, on the employee's departure from the company. A dataset containing 1470 data rows with 14 attributes belonging to the company's employees was selected for the analysis. The method of self-organising Kohonen maps was used, which allow to study the structure of the data and identify hidden patterns, as well as the method of artificial neural networks, which allow to analyse large amounts of data and find hidden relationships that may not be obvious to humans. In the course of the work, the errors of the methods were determined, several experiments with different number of factors were conducted, and the dependence between the number of factors and the magnitude of the error of the algorithms was revealed. For both methods and each experiment, conjugacy tables were obtained, which contain the classification results obtained by the methods. In addition, a correlation analysis was performed to determine the degree of association between the factors and the target variable.

1 Introduction

Humans have always sought to understand how different aspects of their lives relate to each other. This endeavour has arisen from the human need to understand and control the world around them. For example, problems with finances lead people to look for the relationship between income and expenses, the purpose of which is to find those cost items that have a greater impact on the budget.

People seek answers to questions about why one feels the way one feels, why certain events happen, and how one can manage it. Understanding how different aspects of life are connected can help people understand themselves and others better and make more informed decisions. It can also help us to become more effective in achieving our goals.

However, understanding the links between different aspects of life can be a complex and multifaceted process. It requires a lot of research, observation and data analysis. In addition, understanding the connections can lead to new questions that need to be answered.

* Corresponding author: ankoz9@yandex.ru

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).
Nevertheless, the desire to understand how different aspects of life are interconnected is one of the most important human needs that continues to inspire scientists and researchers throughout human history.

In firms, people may seek to find connections between different aspects of the business. For example, people in firms may conduct research and analyse data to identify links between sales and marketing plans, between production and purchasing, between finance and investment, and so on. This helps the firm make better decisions and improve its business strategy.

The corporate environment is the set of conditions in which a company operates. It includes everything related to the work of employees, including company policies, rules and procedures, management system, organisational structure, culture, etc.

The corporate environment has a significant impact on the employees of companies. It determines working conditions, salary levels and career opportunities.

One of the main factors influencing the atmosphere in the team is the leadership style. If a manager is able to create a team, it will have a positive impact on the productivity and motivation of employees.

In addition, the corporate environment affects the emotional state of employees. If the company has a friendly atmosphere, employees will feel more comfortable and confident.

Finally, the corporate environment can influence the level of innovation and development of a company. If a company has a strong corporate culture, it can become a leader in its industry and attract top talent.

2 Materials and methods

Data analysis is the process of extracting information from data and interpreting it to make decisions. It helps to identify hidden patterns and relationships between different variables, which can help to optimise processes.

There are many methods of analysing data, each with its own advantages and disadvantages. For example, there are:

1. Correlation analysis which helps to determine the degree of relationship between two variables.

2. Regression analysis, which is used to identify the relationship between one variable and several other variables.

3. Factor analysis is used to reduce the number of variables by identifying the main factors that determine their relationships.

4. Cluster analysis is used to group objects based on their similarities or differences.

5. Machine learning algorithms automate the process of data analysis and make decisions based on the results.

6. Data visualisation helps to present data in a visual way and reveal hidden patterns.

7. Statistical analysis involves calculating statistical measures such as mean, standard deviation, variance, etc. and interpreting them.

8. Natural language processing is used to analyse textual data and extract useful information from it.

In this paper, cluster data analysis, namely Kohonen’s self-organising map method was used. Data clustering is a data analysis technique that divides a set of points into groups or clusters, so that points belonging to one cluster are more similar to each other than to points from other clusters. This method can be used to identify patterns and relationships in data, as well as to classify objects or predict future events.
Kohonen maps is a data clustering method that is used to group data points based on their proximity to each other. It is based on the idea that data points that are close to each other should be located next to each other on a Kohonen map. Kohonen maps can be used to analyse data in various fields such as medicine, economics, sociology, etc. [11, 12].

The second tool for analysing data was neural networks, which are machine learning algorithms that mimic the workings of the human brain. These algorithms can be used to identify correlations between different factors, as well as to determine the most significant factors affecting the result.

Some of the advantages of neural networks include the following.

1. Adaptability. Neural networks can adapt to different input data and tasks, which allows them to train on a large amount of data and generate more accurate results.
2. Generalisation. Neural networks are able to train on some data and apply the learnt knowledge to other data, thus avoiding overtraining.
3. Processing unstructured data. Neural networks work well with text, images and other unstructured data, making them useful for analysing large amounts of data.
4. Learning without a teacher. Neural networks can independently identify the most important features and relationships in the data, which simplifies learning and improves the accuracy of results.
5. High speed. Neural networks are very fast, which allows them to process large amounts of information in a short amount of time.
6. Flexibility. Neural networks allow you to customise them for specific tasks, maximising efficiency and accuracy [13-15].

A dataset containing company information regarding the departure of employees was selected for the work. The set includes 1470 records and 14 attributes:

- Age - contains the value of the employee's age in years;
- Attrition - informs about the employee's presence in the company - whether he/she has left or is working;
- Department - stores information about the employee's affiliation with a particular department;
- Distance from work to home - estimated distance from workplace to employee's home;
- Education - contains information about the employee's education level;
- Education Field - stores information about the field of education to which the employee's speciality belongs;
- Environmental Satisfaction - employee's assessment of the environment;
- Job satisfaction - employee's evaluation of his/her attitude towards the work process;
- Marital Status - provides information about the employee's marital status;
- Monthly Income - provides information about the employee's available monthly salary;
- Num Companies Worked - provides information on the number of companies the employee worked for before he/she started working for this company;
- Work Life Balance - employee's assessment of the ratio of time spent on work time to personal life;
- Years At Company - stores the number of years the employee has been with the company.

3 Results
The attributes containing string values have been changed. Thus, in the Attrition attribute the No string was assigned the value "0", and the Yes string -"1". The Department attribute began to have values "0", "1", "2", which replaced the Sales, Research & Development, Other strings, respectively.

For the Education attribute, the values "1", "2", "3", "4", "5" were already assigned to the rows Other, College, Bachelor's, Master's, Doctor, respectively. The values of the attributes Environment Satisfaction, Job Satisfaction, Work Life Balance have the following decoding: 1 - low, 2 - middle, 3 - good and 4 - high score. In the Marital Status attribute the lines Single, Married, Other have the values "0", "1", "2" respectively.

Deductor Studio software for data mining and machine learning was used for data analysis. It is designed to process large amounts of data and search for hidden patterns. Deductor allows you to create data models and train them on different data sets. Also, it provides the ability to apply these models for prediction and decision making in different fields such as business, medicine and science [16, 17].

To configure data analysis methods, Deductor has a tool called "Processing Wizard", which is used to set the configurations of the methods to be used. For example, it was used to set the parameters of the Kohonen map algorithm - data columns were assigned, e.g. input and output fields were defined, relations and characteristics of training and test sets were set, dimensionality of maps was set, etc. The work resulted in the Kohonen maps themselves, which are shown in Figure 2, and a contiguity table, which is a two-dimensional array whose rows correspond to the actual classification values and the values in the columns are obtained by running the algorithm. This table is depicted in Figure 1.

**Fig. 1.** Contingency table, Kohonen map method

<table>
<thead>
<tr>
<th></th>
<th>In fact</th>
<th>0</th>
<th>1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td>1211</td>
<td>22</td>
<td>1233</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>197</td>
<td>40</td>
<td>237</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1408</td>
<td>62</td>
<td>1470</td>
</tr>
</tbody>
</table>

The data in the contiguity table reports the following. There are a total of 1233 departing employees and 237 employed employees. The algorithm assigned 1408 items to the "retired" group, 1211 of which were identified correctly, while 197 people should have been identified in the "employed" group. At the same time, 62 people were assigned to the "employed" group, 40 of which were identified correctly and 22 were incorrectly placed in this group.

Each map is allocated to a specific attribute. Maps consist of a set of points, each of which represents a different data element. The distance between the points reflects the degree to which they are similar or different from each other [18].

The programme also produced a correlation table, which is shown in Figure 4. The table contains values that show the relationship between variables, in this case between the Attrition variable and the others. Interpretation of the values of the correlation table contains Table 1.

A negative relationship indicates an inverse relationship and a positive relationship indicates a direct relationship. The table shows that the variables Education Field, Distance for home and Num Companies Worked have a direct weak relationship with Attrition, while the other variables have an inverse weak relationship [4, 19].

Using the algorithm, the data was divided into three clusters. A cluster is a grouping of several items that share a common theme, idea, or goal. In computer science, clusters are used to group similar data [20].
Table 1. Interpretation of correlation coefficient values

<table>
<thead>
<tr>
<th>Correlation coefficient value interval</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.00 ≤ r &lt; -0.70</td>
<td>strong negative</td>
</tr>
<tr>
<td>-0.69 ≤ r &lt; -0.30</td>
<td>moderate negative</td>
</tr>
<tr>
<td>-0.29 ≤ r &lt; -0.01</td>
<td>weak negative</td>
</tr>
<tr>
<td>0.01 &lt; r ≤ 0.29</td>
<td>weak positive</td>
</tr>
<tr>
<td>0.30 &lt; r ≤ 0.69</td>
<td>moderate positive</td>
</tr>
<tr>
<td>0.70 &lt; r ≤ 1.00</td>
<td>strong positive</td>
</tr>
</tbody>
</table>

The zero cluster included 595 employees, the first cluster included 438 employees, and the second cluster included 437 employees. For the zero and first clusters all factors were significant. And for the second cluster the factors Num Companies Worked and Distance from home turned out to be insignificant. It should be noted that for the second cluster one of the least significant factors was Years At Company, but since its significance level is 49.6%, which is not so small, we can conclude that it is also a significant factor. In general, all factors are significant. Figure 3 depicts the parameters of the clusters.

If all the classification factors are significant, it may indicate that each of them has a certain influence on the classification result. Changing the value of one factor can affect the classification result.

The second method of finding dependencies and making classification used in this paper was artificial neural networks. Neural networks are an important tool for data classification. They are machine learning algorithms consisting of many interconnected neurons.
A neuron takes data as input and performs calculations based on this data, passing the information to other neurons in the network.

Fig. 3. Parameters of clusters

The parameters of a neural network are of great importance for its performance. Each of them can affect the quality and accuracy of the model, as well as the speed of its training.

Using the same tool used in the first method, the neural network was configured: input and output fields were set, the test and training sets were configured, the number of hidden layers and neurons in each layer was specified, the activation function was set, and so on.

<table>
<thead>
<tr>
<th>Fields</th>
<th>Indicators</th>
<th>Clusters</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>12 Years At Company</td>
<td>Significance</td>
<td>595 (40.5%)</td>
<td>100.0%</td>
<td>100.0%</td>
<td>49.6%</td>
<td>100.0%</td>
</tr>
<tr>
<td>12 Num Companies Worked</td>
<td>Significance</td>
<td>438 (29.8%)</td>
<td>100.0%</td>
<td>100.0%</td>
<td>13.3%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Monthly Income</td>
<td>Significance</td>
<td>437 (29.7%)</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>12 Marital Status</td>
<td>Significance</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Education</td>
<td>Significance</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Age</td>
<td>Significance</td>
<td>100.0%</td>
<td>100.0%</td>
<td></td>
<td></td>
<td>100.0%</td>
</tr>
<tr>
<td>Environment Satisfaction</td>
<td>Significance</td>
<td>100.0%</td>
<td>100.0%</td>
<td></td>
<td></td>
<td>100.0%</td>
</tr>
<tr>
<td>Education Field</td>
<td>Significance</td>
<td>97.1%</td>
<td>100.0%</td>
<td></td>
<td></td>
<td>100.0%</td>
</tr>
<tr>
<td>Job Satisfaction</td>
<td>Significance</td>
<td>78.6%</td>
<td>99.3%</td>
<td></td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>12 Department</td>
<td>Significance</td>
<td>99.8%</td>
<td>93.4%</td>
<td>92.8%</td>
<td>99.7%</td>
<td></td>
</tr>
<tr>
<td>Distance from home</td>
<td>Significance</td>
<td>92.6%</td>
<td>96.5%</td>
<td>14.0%</td>
<td>99.4%</td>
<td></td>
</tr>
<tr>
<td>Work Life Balance</td>
<td>Significance</td>
<td>70.1%</td>
<td>80.8%</td>
<td>85.7%</td>
<td>75.8%</td>
<td></td>
</tr>
</tbody>
</table>
Figure 6 shows the graph of the first neural network, which contains only one hidden layer with two neurons. Figure 5 shows the contiguity table obtained as a result of the algorithm. It can be seen that 1313 people were counted in the group of departed employees, 127 of which are actually employed. The algorithm assigned 157 people to the number of employed, 47 of whom are not employed. The error of this algorithm was 11.8%.

A neural network graph is a graphical representation of the neural network architecture, which consists of layers of neurons and connections between them. Each layer contains a certain number of neurons that process input data and transmit it further along the network. The connections between neurons determine the direction of information transmission and its intensity. The graph of a neural network allows to visualise the structure of the network and its operation, which simplifies understanding and optimisation of the model [13, 18].

For the second neural network the parameters were changed—the hidden layers became three and the number of neurons on them was increased. Figure 7 shows the obtained adjacency table. Figure 9 shows the graph of this neural network. The algorithm assigned 1328 people to the departed employees, 107 of which have the status of employed. And 142 people were assigned to the employed group, 12 of which are listed as retired. The overall error was 8.1%, which is slightly less than the previous version of the network.

According to the correlation table presented in Figure 4, we can see that three factors have a direct relationship with the target variable: Education Field, Distance from home, Number of Companies Worked. For the next work of the algorithm only these factors were taken. The resulting contiguity table is shown in Figure 8 and the graph of the neural network is shown in Figure 10. An experiment was conducted using two attributes—Education Field, Distance from home and also in a separate experiment the attributes with inverse relationship were tested. The results of this work are recorded in Table 2.
Then all factors were divided according to two criteria. The first criterion - factors not related to the place of work: Age, Education, Education Field, Marital Status, Num Companies Worked. The second criterion is factors that have a relationship with the place of work: Distance from home, Department, Environment Satisfaction, Job Satisfaction, Monthly Income, Work Life Balance, Years At Company. The results of these two experiments are also recorded in Table 2.
Fig. 9. Neural network graph (21x10x5x2x1)

Fig. 10. Neural network graph (5x10x5x2x1)

Table 2. Experimental results
<table>
<thead>
<tr>
<th>№</th>
<th>Experiment</th>
<th>Number of factors</th>
<th>Number of errors</th>
<th>Share of the total number (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>All factors</td>
<td>12</td>
<td>119</td>
<td>8.1</td>
</tr>
<tr>
<td>2</td>
<td>Three factors with a positive correlation coefficient</td>
<td>3</td>
<td>222</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>Two factors with positive correlation coefficient</td>
<td>2</td>
<td>224</td>
<td>15.2</td>
</tr>
<tr>
<td>4</td>
<td>Factors with negative correlation coefficient</td>
<td>9</td>
<td>144</td>
<td>9.7</td>
</tr>
<tr>
<td>5</td>
<td>Factors of the first criterion</td>
<td>5</td>
<td>207</td>
<td>14.1</td>
</tr>
<tr>
<td>6</td>
<td>Factors of the second criterion</td>
<td>7</td>
<td>159</td>
<td>10.8</td>
</tr>
</tbody>
</table>

Figure 11 shows a graph of the values obtained during the work. It can be seen that the more factors are used in the work of algorithms, the lower the number of errors made. But it does not mean that the involvement of all attributes will be the right solution, because the collection of data on each factor requires different resources, besides, the more data, the more time and hardware resources are required to analyse them.

4 Conclusion

Intelligent data analysis can help solve many tasks, such as: building a predictive model to identify future trends and events, analysing large volumes of data to identify hidden patterns and correlations, classifying and clustering data to group similar objects and identify groups, identifying the most important factors affecting results, detecting anomalies and errors in data, automating routine tasks and optimising business processes, developing recommendations for selecting products and services based on the results of data analysis, and identifying the most important factors affecting results.

Data analysis can be a time-consuming process, especially if there is a lot of data or its structure is complex. However, as technology advances and machine learning algorithms improve, this process is becoming more and more efficient and automated [16].
This paper presents an analysis of a dataset that contains 1470 records of people who work and have worked in a particular company. The analysis was performed using Kohonen maps and artificial neural networks. The first method categorized the data into three clusters based on the importance of the attributes. The result of the method revealed that all attributes were significant almost equally. The error of this method was 11.9%.

The second tool of analysis was neural networks. The error of this method was 8.1%, which is slightly better than the first result. Three factors with positive correlation coefficients, two factors with positive correlation coefficients and a group of attributes with negative correlation coefficients were selected for experimentation using the correlation table. The experiment conducted on the group of attributes with negative correlation coefficients resulted in a better result of 9.7% than the experiment on the groups of factors with positive correlation coefficients of 15% and 15.2%.

The attributes were further categorized into two criteria - factors not related to workplace and factors having a relationship with workplace. The error of the algorithm with the first criterion group is 14.1% and with the second criterion group is 10.8%.

According to the results of the two methods, it was not possible to explicitly identify the specific factors affecting the status of the employee. But it can be concluded that the greater the number of factors used, the error of the algorithms becomes smaller. But it is also worth noting that an increase in the number of attributes used leads to an increase in the costs required for analysis.

References


7. I. I. Bosikov, N. Martyushev, R. Klyuev, I. A. Savchenko, V. V. Kukartsev, V. A. Kukartsev, Y. A. Tynchenko, Modeling and complex analysis of the topology parameters of ventilation networks when ensuring fire safety while developing coal and gas deposits, Fire, 6(3), 95 (2023) doi:10.3390/fire6030095


14. V. Kukartsev, A. Mikhalev, A. Stashkevich, K. Moiseeva, Analysis of Data in solving the problem of reducing the accident rate through the use of special means on public roads, 2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), IEEE, 1-4 (2022) doi:10.1109/IEMTRONICS55184.2022.979584


