ANALYSIS OF THE MOST EFFECTIVE DATA MINING TOOLS FOR STATISTICAL ANALYSIS OF THE NAVIGATOR MODEL

Masonkova M., Dyagileva O., Nahrybelnyi Ya., Nosov P.
Kherson State Maritime Academy, Ukraine

Abstract: Throughout the career a navigator is in search of the most optimal trajectory of the educational process and career growth in order to develop all the necessary professional skills. Stakeholders (crewing, shipping companies), in turn, are constantly in search of employees whose professional competence would best meet the requirements of the maritime industry of today. That is why the purpose of this study is to analyze the existing methods and options for the analysis of data mining based on information received from navigators and about navigators' models, which in the future could be used to identify the fundamental factors affecting the level of professional competencies in accordance with the demands of the labor market and, accordingly, the level of their competitiveness.

Novelty. Despite many studies (scientific, sociological, economic, etc.), there is still no single list of criteria for assessing the professional growth of a navigator that would meet all the necessary requirements of stakeholders and would consider the factor of constant dynamic development of the maritime industry.

Keywords: navigator, navigators' model, data mining, data mining tools, analysis, maritime industry.

Introduction. Data mining is the process of discovering patterns in large datasets [1-3]. It consists of a range of methods, from statistical methods through machine learning algorithms to databases. Data mining also includes issues related to data pre-analysis (data pre-processing), model building, data inference, and data visualization. Unlike typical data analysis, which focuses on testing models and hypotheses, data mining uses machine learning algorithms and statistical models to uncover new knowledge and discover hidden patterns in large datasets. Data mining is designed to explore a data set to identify groups of records (cluster analysis), anomalies (outliers), and data correlations. These tasks are usually difficult or impossible to accomplish with traditional methods [4].

Related work. Data mining uses a number of methods, techniques and tools. These are traditional mathematical methods, including statistical methods (mean, standard deviation, etc.), data visualization methods (charts), artificial intelligence tools: neural networks, machine learning, evolutionary methods, fuzzy logic and approximate sets [5]. These data mining methods and tools are presented in specialized literature and implemented in a number of computer applications [6,7-15]. In 2020, scientists from Poland conducted a study on the use of data mining tools to analyze the behavior model of a navigator [16]. Surveys of already employed navigators were taken as the basis for collecting information.

At the first stage, simulation tests of the behavior of navigators in collision situations (AIS data) were carried out. This stage consisted of pre-processing and visualization of the recorded data used to assess the behavior of the navigators by experts. The data obtained in this way were used at the second stage of creating a model (as classifiers) for identifying the navigator's profile. Statistical and artificial intelligence methods are used for data integration, fusion and aggregation, preprocessing, identification and model validation.

Data Collection and Research Methodology. Next, it is necessary to use special software. There are a number of specialized data analysis systems, including:

- PolyAnalyst – it is a software platform for the visual development of scripts for data and text analysis, as well as the construction of interactive reports that do not require programming skills for analytics.

- Anaconda – it is an open source data analysis package management platform (for Python and R languages).

- Gephi – it is an open source data visualization and exploration software specializing in graphs and most species networks.

- IBM SPSS Statistics – it is an analytical software that allows you to perform advanced statistical analysis of business data, covering the solution of all tasks from planning and data collection to direct analysis and business reporting.

The authors of this study propose to use IBM SPSS Statistics for data mining in the future. SPSS Statistics software platform from IBM is designed for statistical data analysis, allowing you to extract useful insights from your data. IBM SPSS Statistics is used by many companies, research centers and independent analytical agencies to solve their own specific business problems, providing quality solutions.

How to work with IBM SPSS Statistics: all the received data are entered into the database, which, in turn, forms data sets for work. For example: 1st data set - a list of professional competencies of a navigator, 2nd data set - technical characteristics of ships, 3d data set - employment statistics for navigators of different ranks on ships of different categories. In other words, the resulting data array must be systematized. For more accurate results, the use of numerical data is recommended. The next stage is the visualization and schematic representation of the obtained data.

There are many criteria and variables that are often used in data mining. For convenience, it is proposed to create aggregated data, for example: acquired professional competencies, work experience in the specialty, position held, salary level, all these indicators can be combined into one aggregated, for example, “professional level”. In this way, it is possible to assemble data of any nature.

In order to create professional profiles of navigators, it is proposed to use cluster analysis (partitioning a set of elements into relatively homogeneous groups or clusters). An important advantage of cluster analysis, along with other classical modeling methods, is the ability to classify objects not by one parameter, but by a certain set of features. Such studies can be carried out for a variety of initial data of an almost arbitrary nature. This is of great practical importance in the presence of heterogeneous indicators that make it difficult to use traditional mathematical approaches. Cluster analysis allows you to analyze fairly large amounts of data and drastically reduce, compress large amounts of information, make them compact and visual. The solution to the classification problem is to assign each of the data objects to one (or several) of the predefined classes and, ultimately, to build with one of the methods for classifying a data model that determines the division of a set of data objects into classes.

To identify the individual characteristics of navigators that determine their likelihood of employment and compliance with the requirements of stakeholders, it is proposed to use a scoring model (weighted scorecard). It is a technique for weighing certain decisions. This mathematical model is used to prioritize strategies and decisions by assigning a numeric value. Scoring models are built using statistical methods (Chi square test or ROC-curve) to the studied data array (nominal and numerical features) in order to assess the risks of an event and find significant factors influencing the predicted probability of its occurrence.

The use of several different tools to predict the further professional development of a navigator and his further employment allows us to identify latent relationships and confirm

preliminary conclusions. For applied research, it is planned to use the following Data Mining tools (Fig. 1).

In the studied research [17], Decision Tree was chosen as the most optimal method of data mining. Also, factor and cluster analyze are quite popular. The results of factor analysis in the study "Exploring technical and non-technical competencies of navigators for autonomous shipping" can become part of the basis for subsequent research related to the identification of the main parameters of the navigator model that are significant for stakeholders.

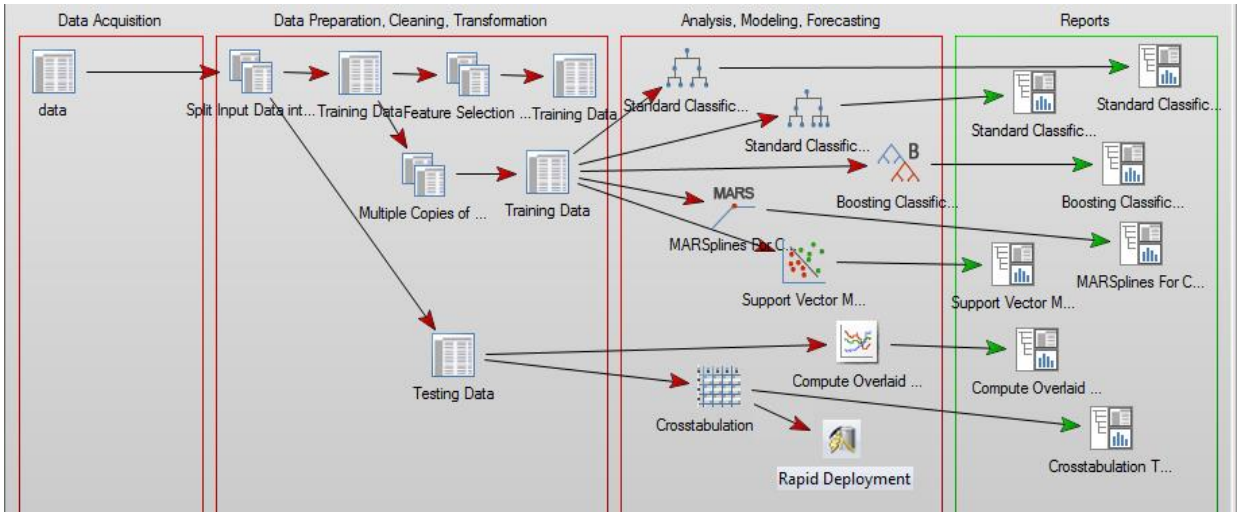

Fig. 1. Data mining project window

Previously, the authors of this study studied the use of graphs in modeling the navigator model [18-21]. This approach allows to consider the state space, which is also an important point when building a navigator model. For example, the authors built a graph model of a navigator.

Conclusions. An analysis of existing studies aimed at identifying the navigator model allows us to build a clear algorithm of the actions we need, so that in the shortest possible time, considering the time for collecting and processing empirical data, we can obtain results that could work in practice. One of the main conclusions is that qualitative research is impossible without the use of data mining tools. As the fundamental methods of data analysis, we have chosen factor and cluster analysis, Decision Tree and the method of using graphs. Together, these methods will make it possible to build a navigator's model, considering all quantitative parameters and the possibility of a clear graphical representation of both the model itself and the processes occurring during its existence. It is more efficient to experiment with different methods during data mining or modeling than to rely on one of these methods. Various tools help to understand the problem as a whole or verify preliminary conclusions. In view of the complexity of the problem under study, it is proposed to conduct empirical studies using several different methods for the same real statistical data.

References

1. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2009.

2. Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques, 3rd ed.; Morgan Kaufmann: Burlington, MA, USA, 2012.

3. Fayyad U., Piatetsky-Shapiro G., Smyth P. From Data Mining to Knowledge Discovery in Databases.

4. Kantardzic M. Data Mining: Concepts, Models, Methods, and Algorithms; John Wiley & Sons: Hoboken, NJ, USA, 2003.

5. Hofmann M., Klinkenberg R. RapidMiner: Data Mining Use Cases and Business Analytics Applications. In Chapman & Hall/CRC Data Mining and Knowledge Discovery Series; CRC Press: Boca Raton, FL, USA, 2013. J. Mar. Sci. Eng. 2020, 8, 50 16 of 16.

6. Zhang Y.A. Comprehensive Survey on Particle Swarm Optimization Algorithm and Its Applications. Math. Probl. Eng. 2015, 2015, 931256. [CrossRef].

7. Anand D. Feature extraction for collaborative filtering: A genetic programming approach. Int. J. Comput. Sci. Issues 2012, 9, 348.

8. Plokhikh, V., Popovych, I., Zavatska, N., Losiyevska, O., Zinchenko, S., Nosov, P., & Aleksieieva, M. (2021). Time Synthesis in Organization of Sensorimotor Action. BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 12(4), 164-188. https://doi.org/10.18662/brain/12.4/243.

9. Зинченко С.Н., Носов П.С., Грошева О.А., Маменко П.П., Матейчук В.Н. Управление судном в условиях внешних воздействий. Materials of the XI "Modern information technologies in transport, MINTT-2019" May 28-30, 2019 Kherson, Ukraine. С 177-178.

10. Зинченко С.Н., Носов П.С., Грошева О.А., Маменко П.П., Матейчук В.Н. Избыточность по управлению как количественная мера маневренности судна. Materials of the XI "Modern information technologies in transport, MINTT-2019" May 28-30, 2019 Kherson, Ukraine. С 97-99.

11. Nosov P.S., Zinchenko S.M., Ben A.P., Nahrybelnyi Ya. A., Dudchenko O.M. MODELS OF DECISION MAKING BY A NAVIGATOR UNDER IMPLICIT AGREEMENTS WITH COLREG RULES // Науковий вісник Херсонської державної морської академії: науковий журнал. – Херсон: Херсонська державна морська академія, 2019. – № 1 (20). – С. 31-38.

12. Nosov P., Cherniavskyi V., Zinchenko S., Popovych I., Prokopchuk Y., Safonov M. Identification of distortion of the navigator's time in model experiment // Bulletin of University of Karaganda. Instrument and experimental techniques, 2020. - № 4(100). P. 57-70. DOI: 10.31489/2020Ph4/57-70.

13. Pietrzykowski Z., Wielgosz M., Breitsprecher M. Navigators' behavior analysis using data mining. J. Mar. Sci. Eng. 2020, 8, 50.

14. Nosov P., Krapyvko G., Ben A., Safonov M., Zinchenko S. Disabling the dynamic positioning of the vessel as a cause of the negative influence of human factor in maritime transport. МНПК пам'яті професорів Фоміна Ю. Я. і Семенова В. С. (FS - 2019), 24 – 28 квітня 2019, Одеса – Стамбул – Одеса. Pages 309-315.

15. Zinchenko, S. M., Ben, A. P., Nosov, P. S., Popovych, I. S., Mamenko, P. P., & Mateichuk, V. M. (2020). Improving the accuracy and reliability of automatic vessel moution control system. Radio Electronics, Computer Science, Control, (2), 183–195. https://doi.org/10.15588/1607-3274-2020-2-19.

16. Amit Sharma & Tae-eun Kim (2022) Exploring technical and non-technical competencies of navigators for autonomous shipping, Maritime Policy & Management, 49:6, 831-849, DOI: 10.1080/03088839.2021.1914874

17. Masonkova M., Diahyleva O.S., Ben A.P., Nosov P.S. FORMAL APPROACHES FOR IDENTIFICATION STATE-SPACE NAVIGATOR'S MODELS ON GRAPHS / Materials of the XIV international scientific and practical conference "Modern information technologies in transport, MINTT-2022", 2022 Kherson, Ukraine.

18. Cherniavskyi V., Diahyleva O., Masonkova M., Nosov P. FORMAL-LOGICAL APPROACHES TO BUILDING INFORMATION MODEL OF NAVIGATOR IN THE FORM OF A DYNAMIC TRAJECTORY / MPP&O-2022, Одеса – Стамбул – Одеса. С. 363-368.

19. Mariia Masonkova, Olena Dyagileva, Pavlo Nosov. Development of the identification system of cadets' qualification characteristics regarding stakeholder // Сучасні енергетичні установки на транспорті і технології та обладнання для їх обслуговування (СЕУТТО 2021). С. 280-282.

20. Nosov, P. S., Popovych, I. S., Cherniavskyi, V. V., Zinchenko, S. M., Prokopchuk, Y. A., & Makarchuk, D. V. (2020). Automated identification of an operator anticipation on marine transport. Radio Electronics, Computer Science, Control, (3), 158–172. https://doi.org/10.15588/1607-3274-2020-3-15.

21. Zinchenko S., Nosov P., Mateichuk V., Mamenko P., Popovych I., Grosheva O. Automatic collision avoidance system with many targets, including maneuvering ones // Bulletin of University of Karaganda. Technical Physics, 2019. - № 4(96). P. 69-79. https://doi.org/10.31489/2019Ph4/69-79.